Chapter Three

Multiple Regression Analysis: Estimation

In Chapter 2, we learned how to use simple regression analysis to explain a dependent variable, y, as a function of a single independent variable, x. The primary drawback in using simple regression analysis for empirical work is that it is very difficult to draw ceteris paribus conclusions about how x affects y: the key assumption, SLR.3—that all other factors affecting y are uncorrelated with x—is often unrealistic.

Multiple regression analysis is more amenable to ceteris paribus analysis because it allows us to *explicitly* control for many other factors which simultaneously affect the dependent variable. This is important both for testing economic theories and for evaluating policy effects when we must rely on nonexperimental data. Because multiple regression models can accommodate many explanatory variables that may be correlated, we can hope to infer causality in cases where simple regression analysis would be misleading.

Naturally, if we add more factors to our model that are useful for explaining y, then more of the variation in y can be explained. Thus, multiple regression analysis can be used to build better models for predicting the dependent variable.

An additional advantage of multiple regression analysis is that it can incorporate fairly general functional form relationships. In the simple regression model, only one function of a single explanatory variable can appear in the equation. As we will see, the multiple regression model allows for much more flexibility.

Section 3.1 formally introduces the multiple regression model and further discusses the advantages of multiple regression over simple regression. In Section 3.2, we demonstrate how to estimate the parameters in the multiple regression model using the method of ordinary least squares. In Sections 3.3, 3.4, and 3.5, we describe various statistical properties of the OLS estimators, including unbiasedness and efficiency.

The multiple regression model is still the most widely used vehicle for empirical analysis in economics and other social sciences. Likewise, the method of ordinary least squares is popularly used for estimating the parameters of the multiple regression model.

3.1 MOTIVATION FOR MULTIPLE REGRESSION

The Model with Two Independent Variables

We begin with some simple examples to show how multiple regression analysis can be used to solve problems that cannot be solved by simple regression.

The first example is a simple variation of the wage equation introduced in Chapter 2 for obtaining the effect of education on hourly wage:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u, \tag{3.1}$$

where *exper* is years of labor market experience. Thus, *wage* is determined by the two explanatory or independent variables, education and experience, and by other unobserved factors, which are contained in u. We are still primarily interested in the effect of *educ* on *wage*, holding fixed all other factors affecting *wage*; that is, we are interested in the parameter β_1.

Compared with a simple regression analysis relating *wage* to *educ*, equation (3.1) effectively takes *exper* out of the error term and puts it explicitly in the equation. Because *exper* appears in the equation, its coefficient, β_2, measures the ceteris paribus effect of *exper* on *wage*, which is also of some interest.

Not surprisingly, just as with simple regression, we will have to make assumptions about how u in (3.1) is related to the independent variables, *educ* and *exper*. However, as we will see in Section 3.2, there is one thing of which we can be confident: since (3.1) contains experience explicitly, we will be able to measure the effect of education on wage, holding experience fixed. In a simple regression analysis—which puts *exper* in the error term—we would have to assume that experience is uncorrelated with education, a tenuous assumption.

As a second example, consider the problem of explaining the effect of per student spending (*expend*) on the average standardized test score (*avgscore*) at the high school level. Suppose that the average test score depends on funding, average family income (*avginc*), and other unobservables:

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u. \tag{3.2}$$

The coefficient of interest for policy purposes is β_1, the ceteris paribus effect of *expend* on *avgscore*. By including *avginc* explicitly in the model, we are able to control for its effect on *avgscore*. This is likely to be important because average family income tends to be correlated with per student spending: spending levels are often determined by both property and local income taxes. In simple regression analysis, *avginc* would be included in the error term, which would likely be correlated with *expend*, causing the OLS estimator of β_1 in the two-variable model to be biased.

In the two previous similar examples, we have shown how observable factors other than the variable of primary interest [*educ* in equation (3.1), *expend* in equation (3.2)] can be included in a regression model. Generally, we can write a model with two independent variables as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.3}$$

where β_0 is the intercept, β_1 measures the change in y with respect to x_1, holding other factors fixed, and β_2 measures the change in y with respect to x_2, holding other factors fixed.

Multiple regression analysis is also useful for generalizing functional relationships between variables. As an example, suppose family consumption (*cons*) is a quadratic function of family income (*inc*):

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u, \tag{3.4}$$

where u contains other factors affecting consumption. In this model, consumption depends on only one observed factor, income; so it might seem that it can be handled in a simple regression framework. But the model falls outside simple regression because it contains two functions of income, *inc* and inc^2 (and therefore three parameters, β_0, β_1, and β_2). Nevertheless, the consumption function is easily written as a regression model with two independent variables by letting $x_1 = inc$ and $x_2 = inc^2$.

Mechanically, there will be *no* difference in using the method of ordinary least squares (introduced in Section 3.2) to estimate equations as different as (3.1) and (3.4). Each equation can be written as (3.3), which is all that matters for computation. There is, however, an important difference in how one *interprets* the parameters. In equation (3.1), β_1 is the ceteris paribus effect of *educ* on *wage*. The parameter β_1 has no such interpretation in (3.4). In other words, it makes no sense to measure the effect of *inc* on *cons* while holding inc^2 fixed, because if *inc* changes, then so must inc^2! Instead, the change in consumption with respect to the change in income—the marginal propensity to consume—is approximated by

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc.$$

See Appendix A for the calculus needed to derive this equation. In other words, the marginal effect of income on consumption depends on β_2 as well as on β_1 and the level of income. This example shows that, in any particular application, the definition of the independent variables are crucial. But for the theoretical development of multiple regression, we can be vague about such details. We will study examples like this more completely in Chapter 6.

In the model with two independent variables, the key assumption about how u is related to x_1 and x_2 is

$$\mathrm{E}(u|x_1,x_2) = 0. \tag{3.5}$$

The interpretation of condition (3.5) is similar to the interpretation of Assumption SLR.3 for simple regression analysis. It means that, for any values of x_1 and x_2 in the population, the average unobservable is equal to zero. As with simple regression, the important part of the assumption is that the expected value of u is the same for all combinations of x_1 and x_2; that this common value is zero is no assumption at all as long as the intercept β_0 is included in the model (see Section 2.1).

How can we interpret the zero conditional mean assumption in the previous examples? In equation (3.1), the assumption is $\mathrm{E}(u|educ,exper) = 0$. This implies that other factors affecting *wage* are not related on average to *educ* and *exper*. Therefore, if we think innate ability is part of u, then we will need average ability levels to be the same across all combinations of education and experience in the working population. This

may or may not be true, but, as we will see in Section 3.3, this is the question we need to ask in order to determine whether the method of ordinary least squares produces unbiased estimators.

The example measuring student performance [equation (3.2)] is similar to the wage equation. The zero conditional mean assumption is $E(u|expend,avginc) = 0$, which means that other factors affecting test scores—school or student characteristics—are, on average, unrelated to per student funding and average family income.

QUESTION 3.1

A simple model to explain city murder rates (*murdrate*) in terms of the probability of conviction (*prbconv*) and average sentence length (*avgsen*) is

$$murdrate = \beta_0 + \beta_1 prbconv + \beta_2 avgsen + u.$$

What are some factors contained in u? Do you think the key assumption (3.5) is likely to hold?

When applied to the quadratic consumption function in (3.4), the zero conditional mean assumption has a slightly different interpretation. Written literally, equation (3.5) becomes $E(u|inc,inc^2) = 0$. Since inc^2 is known when inc is known, including inc^2 in the expectation is redundant: $E(u|inc,inc^2) = 0$ is the same as $E(u|inc) = 0$. Nothing is wrong with putting inc^2 along with inc in the expectation when stating the assumption, but $E(u|inc) = 0$ is more concise.

The Model with *k* Independent Variables

Once we are in the context of multiple regression, there is no need to stop with two independent variables. Multiple regression analysis allows many observed factors to affect y. In the wage example, we might also include amount of job training, years of tenure with the current employer, measures of ability, and even demographic variables like number of siblings or mother's education. In the school funding example, additional variables might include measures of teacher quality and school size.

The general **multiple linear regression model** (also called the multiple regression model) can be written in the population as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u, \qquad \textbf{(3.6)}$$

where β_0 is the **intercept**, β_1 is the parameter associated with x_1, β_2 is the parameter associated with x_2, and so on. Since there are k independent variables and an intercept, equation (3.6) contains $k + 1$ (unknown) population parameters. For shorthand purposes, we will sometimes refer to the parameters other than the intercept as **slope parameters**, even though this is not always literally what they are. [See equation (3.4), where neither β_1 nor β_2 is itself a slope, but together they determine the slope of the relationship between consumption and income.]

The terminology for multiple regression is similar to that for simple regression and is given in Table 3.1. Just as in simple regression, the variable u is the **error term** or **disturbance**. It contains factors other than $x_1, x_2, \ldots, x_k$ that affect y. No matter how many explanatory variables we include in our model, there will always be factors we cannot include, and these are collectively contained in u.

When applying the general multiple regression model, we must know how to interpret the parameters. We will get plenty of practice now and in subsequent chapters, but

Table 3.1

Terminology for Multiple Regression

y	$x_1, x_2, \ldots, x_k$
Dependent Variable	Independent Variables
Explained Variable	Explanatory Variables
Response Variable	Control Variables
Predicted Variable	Predictor Variables
Regressand	Regressors

it is useful at this point to be reminded of some things we already know. Suppose that CEO salary (*salary*) is related to firm sales and CEO tenure with the firm by

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u. \quad \textbf{(3.7)}$$

This fits into the multiple regression model (with $k = 3$) by defining $y = \log(salary)$, $x_1 = \log(sales)$, $x_2 = ceoten$, and $x_3 = ceoten^2$. As we know from Chapter 2, the parameter β_1 is the (ceteris paribus) *elasticity* of *salary* with respect to *sales*. If $\beta_3 = 0$, then $100\beta_2$ is approximately the ceteris paribus percentage increase in *salary* when *ceoten* increases by one year. When $\beta_3 \neq 0$, the effect of *ceoten* on *salary* is more complicated. We will postpone a detailed treatment of general models with quadratics until Chapter 6.

Equation (3.7) provides an important reminder about multiple regression analysis. The term "linear" in multiple linear regression model means that equation (3.6) is linear in the *parameters*, β_j. Equation (3.7) is an example of a multiple regression model that, while linear in the β_j, is a nonlinear relationship between *salary* and the variables *sales* and *ceoten*. Many applications of multiple linear regression involve nonlinear relationships among the underlying variables.

The key assumption for the general multiple regression model is easy to state in terms of a conditional expectation:

$$E(u|x_1, x_2, \ldots, x_k) = 0. \quad \textbf{(3.8)}$$

At a minimum, equation (3.8) requires that all factors in the unobserved error term be uncorrelated with the explanatory variables. It also means that we have correctly accounted for the functional relationships between the explained and explanatory variables. Any problem that allows u to be correlated with any of the independent variables causes (3.8) to fail. In Section 3.3, we will show that assumption (3.8) implies that OLS is unbiased and will derive the bias that arises when a key variable has been omitted

from the equation. In Chapters 15 and 16, we will study other reasons that might cause (3.8) to fail and show what can be done in cases where it does fail.

3.2 MECHANICS AND INTERPRETATION OF ORDINARY LEAST SQUARES

We now summarize some computational and algebraic features of the method of ordinary least squares as it applies to a particular set of data. We also discuss how to interpret the estimated equation.

Obtaining the OLS Estimates

We first consider estimating the model with two independent variables. The estimated OLS equation is written in a form similar to the simple regression case:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \tag{3.9}$$

where $\hat{\beta}_0$ is the estimate of β_0, $\hat{\beta}_1$ is the estimate of β_1, and $\hat{\beta}_2$ is the estimate of β_2. But how do we obtain $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$? The method of **ordinary least squares** chooses the estimates to minimize the sum of squared residuals. That is, given n observations on y, x_1, and x_2, $\{(x_{i1},x_{i2},y_i)\colon i = 1,2, \ldots, n\}$, the estimates $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are chosen simultaneously to make

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \tag{3.10}$$

as small as possible.

In order to understand what OLS is doing, it is important to master the meaning of the indexing of the independent variables in (3.10). The independent variables have two subscripts here, i followed by either 1 or 2. The i subscript refers to the observation number. Thus, the sum in (3.10) is over all $i = 1$ to n observations. The second index is simply a method of distinguishing between different independent variables. In the example relating *wage* to *educ* and *exper*, $x_{i1} = educ_i$ is education for person i in the sample, and $x_{i2} = exper_i$ is experience for person i. The sum of squared residuals in equation (3.10) is $\sum_{i=1}^{n} (wage_i - \hat{\beta}_0 - \hat{\beta}_1 educ_i - \hat{\beta}_2 exper_i)^2$. In what follows, the i subscript is reserved for indexing the observation number. If we write x_{ij}, then this means the i^{th} observation on the j^{th} independent variable. (Some authors prefer to switch the order of the observation number and the variable number, so that x_{1i} is observation i on variable one. But this is just a matter of notational taste.)

In the general case with k independent variables, we seek estimates $\hat{\beta}_0$, $\hat{\beta}_1$, …, $\hat{\beta}_k$ in the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.11}$$

The OLS estimates, $k + 1$ of them, are chosen to minimize the sum of squared residuals:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2. \tag{3.12}$$

This minimization problem can be solved using multivariable calculus (see Appendix 3A). This leads to $k + 1$ linear equations in $k + 1$ unknowns $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$:

$$\begin{aligned}
&\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \\
&\quad\vdots \\
&\sum_{i=1}^{n} x_{ik}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0.
\end{aligned} \tag{3.13}$$

These are often called the OLS **first order conditions**. As with the simple regression model in Section 2.2, the OLS first order conditions can be motivated by the method of moments: under assumption (3.8), $E(u) = 0$ and $E(x_j u) = 0$, where $j = 1,2, \ldots, k$. The equations in (3.13) are the sample counterparts of these population moments.

For even moderately sized n and k, solving the equations in (3.13) by hand calculations is tedious. Nevertheless, modern computers running standard statistics and econometrics software can solve these equations with large n and k very quickly.

There is only one slight caveat: we must assume that the equations in (3.13) can be solved *uniquely* for the $\hat{\beta}_j$. For now, we just assume this, as it is usually the case in well-specified models. In Section 3.3, we state the assumption needed for unique OLS estimates to exist (see Assumption MLR.4).

As in simple regression analysis, equation (3.11) is called the **OLS regression line**, or the **sample regression function** (**SRF**). We will call $\hat{\beta}_0$ the **OLS intercept estimate** and $\hat{\beta}_1, \ldots, \hat{\beta}_k$ the **OLS slope estimates** (corresponding to the independent variables x_1, $x_2, \ldots, x_k$).

In order to indicate that an OLS regression has been run, we will either write out equation (3.11) with y and $x_1, \ldots, x_k$ replaced by their variable names (such as *wage*, *educ*, and *exper*), or we will say that "we ran an OLS regression of y on $x_1, x_2, \ldots, x_k$" or that "we regressed y on $x_1, x_2, \ldots, x_k$." These are shorthand for saying that the method of ordinary least squares was used to obtain the OLS equation (3.11). Unless explicitly stated otherwise, we always estimate an intercept along with the slopes.

Interpreting the OLS Regression Equation

More important than the details underlying the computation of the $\hat{\beta}_j$ is the *interpretation* of the estimated equation. We begin with the case of two independent variables:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \quad \textbf{(3.14)}$$

The intercept $\hat{\beta}_0$ in equation (3.14) is the predicted value of y when $x_1 = 0$ and $x_2 = 0$. Sometimes setting x_1 and x_2 both equal to zero is an interesting scenario, but in other cases it will not make sense. Nevertheless, the intercept is always needed to obtain a prediction of y from the OLS regression line, as (3.14) makes clear.

The estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ have **partial effect**, or **ceteris paribus**, interpretations. From equation (3.14), we have

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2,$$

so we can obtain the predicted change in y given the changes in x_1 and x_2. (Note how the intercept has nothing to do with the changes in y.) In particular, when x_2 is held fixed, so that $\Delta x_2 = 0$, then

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1,$$

holding x_2 fixed. The key point is that, by including x_2 in our model, we obtain a coefficient on x_1 with a ceteris paribus interpretation. This is why multiple regression analysis is so useful. Similarly,

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2,$$

holding x_1 fixed.

EXAMPLE 3.1
(Determinants of College GPA)

The variables in GPA1.RAW include college grade point average (*colGPA*), high school GPA (*hsGPA*), and achievement test score (*ACT*) for a sample of 141 students from a large university; both college and high school GPAs are on a four-point scale. We obtain the following OLS regression line to predict college GPA from high school GPA and achievement test score:

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT. \quad \textbf{(3.15)}$$

How do we interpret this equation? First, the intercept 1.29 is the predicted college GPA if *hsGPA* and *ACT* are both set as zero. Since no one who attends college has either a zero high school GPA or a zero on the achievement test, the intercept in this equation is not, by itself, meaningful.

More interesting estimates are the slope coefficients on *hsGPA* and *ACT*. As expected, there is a positive partial relationship between *colGPA* and *hsGPA*: holding *ACT* fixed, another point on *hsGPA* is associated with .453 of a point on the college GPA, or almost half a point. In other words, if we choose two students, A and B, and these students have the same ACT score, but the high school GPA of Student A is one point higher than the high school GPA of Student B, then we predict Student A to have a college GPA .453 higher than that of Student B. [This says nothing about any two actual people, but it is our best prediction.]

The sign on *ACT* implies that, while holding *hsGPA* fixed, a change in the ACT score of 10 points—a very large change, since the average score in the sample is about 24 with a standard deviation less than three—affects *colGPA* by less than one-tenth of a point. This is a small effect, and it suggests that, once high school GPA is accounted for, the ACT score is not a strong predictor of college GPA. (Naturally, there are many other factors that contribute to GPA, but here we focus on statistics available for high school students.) Later, after we discuss statistical inference, we will show that not only is the coefficient on ACT practically small, it is also statistically insignificant.

If we focus on a simple regression analysis relating *colGPA* to *ACT* only, we obtain

$$\widehat{colGPA} = 2.40 + .0271\, ACT;$$

thus, the coefficient on *ACT* is almost three times as large as the estimate in (3.15). But this equation does *not* allow us to compare two people with the same high school GPA; it corresponds to a different experiment. We say more about the differences between multiple and simple regression later.

The case with more than two independent variables is similar. The OLS regression line is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.16}$$

Written in terms of changes,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k. \tag{3.17}$$

The coefficient on x_1 measures the change in $\hat{y}$ due to a one-unit increase in x_1, holding all other independent variables fixed. That is,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1, \tag{3.18}$$

holding $x_2, x_3, \ldots, x_k$ fixed. Thus, we have controlled for the variables $x_2, x_3, \ldots, x_k$ when estimating the effect of x_1 on y. The other coefficients have a similar interpretation.

The following is an example with three independent variables.

EXAMPLE 3.2
(Hourly Wage Equation)

Using the 526 observations on workers in WAGE1.RAW, we include *educ* (years of education), *exper* (years of labor market experience), and *tenure* (years with the current employer) in an equation explaining log(*wage*). The estimated equation is

$$\widehat{\log(wage)} = .284 + .092\, educ + .0041\, exper + .022\, tenure. \tag{3.19}$$

As in the simple regression case, the coefficients have a percentage interpretation. The only difference here is that they also have a ceteris paribus interpretation. The coefficient .092

means that, holding *exper* and *tenure* fixed, another year of education is predicted to increase log(*wage*) by .092, which translates into an approximate 9.2 percent [100(.092)] increase in *wage*. Alternatively, if we take two people with the same levels of experience and job tenure, the coefficient on *educ* is the proportionate difference in predicted wage when their education levels differ by one year. This measure of the return to education at least keeps two important productivity factors fixed; whether it is a good estimate of the ceteris paribus return to another year of education requires us to study the statistical properties of OLS (see Section 3.3).

On the Meaning of "Holding Other Factors Fixed" in Multiple Regression

The partial effect interpretation of slope coefficients in multiple regression analysis can cause some confusion, so we attempt to prevent that problem now.

In Example 3.1, we observed that the coefficient on *ACT* measures the predicted difference in *colGPA*, holding *hsGPA* fixed. The power of multiple regression analysis is that it provides this ceteris paribus interpretation even though the data have *not* been collected in a ceteris paribus fashion. In giving the coefficient on *ACT* a partial effect interpretation, it may seem that we actually went out and sampled people with the same high school GPA but possibly with different ACT scores. This is not the case. The data are a random sample from a large university: there were no restrictions placed on the sample values of *hsGPA* or *ACT* in obtaining the data. Rarely do we have the luxury of holding certain variables fixed in obtaining our sample. *If* we could collect a sample of individuals with the same high school GPA, then we could perform a simple regression analysis relating *colGPA* to *ACT*. Multiple regression effectively allows us to mimic this situation without restricting the values of any independent variables.

The power of multiple regression analysis is that it allows us to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed.

Changing More than One Independent Variable Simultaneously

Sometimes we want to change more than one independent variable at the same time to find the resulting effect on the dependent variable. This is easily done using equation (3.17). For example, in equation (3.19), we can obtain the estimated effect on *wage* when an individual stays at the same firm for another year: *exper* (general workforce experience) and *tenure* both increase by one year. The total effect (holding *educ* fixed) is

$$\Delta\widehat{\log(wage)} = .0041\,\Delta exper + .022\,\Delta tenure = .0041 + .022 = .0261,$$

or about 2.6 percent. Since *exper* and *tenure* each increase by one year, we just add the coefficients on *exper* and *tenure* and multiply by 100 to turn the effect into a percent.

OLS Fitted Values and Residuals

After obtaining the OLS regression line (3.11), we can obtain a *fitted* or *predicted value* for each observation. For observation *i*, the fitted value is simply

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}, \tag{3.20}$$

which is just the predicted value obtained by plugging the values of the independent variables for observation i into equation (3.11). We should not forget about the intercept in obtaining the fitted values; otherwise, the answer can be very misleading. As an example, if in (3.15), $hsGPA_i = 3.5$ and $ACT_i = 24$, $\widehat{colGPA}_i = 1.29 + .453(3.5) + .0094(24) = 3.101$ (rounded to three places after the decimal).

QUESTION 3.2

In Example 3.1, the OLS fitted line explaining college GPA in terms of high school GPA and ACT score is

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT.$$

If the average high school GPA is about 3.4 and the average ACT score is about 24.2, what is the average college GPA in the sample?

Normally, the actual value y_i for any observation i will not equal the predicted value, $\hat{y}_i$: OLS minimizes the *average* squared prediction error, which says nothing about the prediction error for any particular observation. The **residual** for observation i is defined just as in the simple regression case,

$$\hat{u}_i = y_i - \hat{y}_i. \tag{3.21}$$

There is a residual for each observation. If $\hat{u}_i > 0$, then $\hat{y}_i$ is below y_i, which means that, for this observation, y_i is underpredicted. If $\hat{u}_i < 0$, then $y_i < \hat{y}_i$, and y_i is overpredicted.

The OLS fitted values and residuals have some important properties that are immediate extensions from the single variable case:

1. The sample average of the residuals is zero.
2. The sample covariance between each independent variable and the OLS residuals is zero. Consequently, the sample covariance between the OLS fitted values and the OLS residuals is zero.
3. The point $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ is always on the OLS regression line: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \ldots + \hat{\beta}_k \bar{x}_k$.

The first two properties are immediate consequences of the set of equations used to obtain the OLS estimates. The first equation in (3.13) says that the sum of the residuals is zero. The remaining equations are of the form $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, which imply that the each independent variable has zero sample covariance with $\hat{u}_i$. Property 3 follows immediately from Property 1.

A "Partialling Out" Interpretation of Multiple Regression

When applying OLS, we do not need to know explicit formulas for the $\hat{\beta}_j$ that solve the system of equations (3.13). Nevertheless, for certain derivations, we do need explicit formulas for the $\hat{\beta}_j$. These formulas also shed further light on the workings of OLS.

Consider again the case with $k = 2$ independent variables, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. For concreteness, we focus on $\hat{\beta}_1$. One way to express $\hat{\beta}_1$ is

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} \hat{r}_{i1} y_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right), \tag{3.22}$$

where the $\hat{r}_{i1}$ are the OLS residuals from a simple regression of x_1 on x_2, using the sample at hand. We regress our first independent variable, x_1, on our second independent variable, x_2, and then obtain the residuals (y plays no role here). Equation (3.22) shows that we can then do a simple regression of y on $\hat{r}_1$ to obtain $\hat{\beta}_1$. (Note that the residuals $\hat{r}_{i1}$ have a zero sample average, and so $\hat{\beta}_1$ is the usual slope estimate from simple regression.)

The representation in equation (3.22) gives another demonstration of $\hat{\beta}_1$'s partial effect interpretation. The residuals $\hat{r}_{i1}$ are the part of x_{i1} that is uncorrelated with x_{i2}. Another way of saying this is that $\hat{r}_{i1}$ is x_{i1} after the effects of x_{i2} have been *partialled out*, or *netted out*. Thus, $\hat{\beta}_1$ measures the sample relationship between y and x_1 after x_2 has been partialled out.

In simple regression analysis, there is no partialling out of other variables because no other variables are included in the regression. Problem 3.17 steps you through the partialling out process using the wage data from Example 3.2. For practical purposes, the important thing is that $\hat{\beta}_1$ in the equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ measures the change in y given a one-unit increase in x_1, holding x_2 fixed.

In the general model with k explanatory variables, $\hat{\beta}_1$ can still be written as in equation (3.22), but the residuals $\hat{r}_{i1}$ come from the regression of x_1 on $x_2, \ldots, x_k$. Thus, $\hat{\beta}_1$ measures the effect of x_1 on y after $x_2, \ldots, x_k$ have been partialled or netted out.

Comparison of Simple and Multiple Regression Estimates

Two special cases exist in which the simple regression of y on x_1 will produce the *same* OLS estimate on x_1 as the regression of y on x_1 and x_2. To be more precise, write the simple regression of y on x_1 as $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$ and write the multiple regression as $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. We know that the simple regression coefficient $\tilde{\beta}_1$ does not usually equal the multiple regression coefficient $\hat{\beta}_1$. There are two distinct cases where $\tilde{\beta}_1$ and $\hat{\beta}_1$ are identical:

1. The partial effect of x_2 on y is zero in the sample. That is, $\hat{\beta}_2 = 0$.
2. x_1 and x_2 are uncorrelated in the sample.

The first assertion can be proven by looking at two of the equations used to determine $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$: $\sum_{i=1}^{n} x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$. Setting $\hat{\beta}_2 = 0$ gives the same intercept and slope as does the regression of y on x_1.

The second assertion follows from equation (3.22). If x_1 and x_2 are uncorrelated in the sample, then regressing x_1 on x_2 results in no partialling out, and so the simple regression of y on x_1 and the multiple regression of y on x_1 and x_2 produce identical estimates on x_1.

Even though simple and multiple regression estimates are almost never identical, we can use the previous characterizations to explain why they might be either very different or quite similar. For example, if $\hat{\beta}_2$ is small, we might expect the simple and mul-

tiple regression estimates of β_1 to be similar. In Example 3.1, the sample correlation between *hsGPA* and *ACT* is about 0.346, which is a nontrivial correlation. But the coefficient on *ACT* is fairly little. It is not suprising to find that the simple regression of *colGPA* on *hsGPA* produces a slope estimate of .482, which is not much different from the estimate .453 in (3.15).

EXAMPLE 3.3
(Participation in 401(k) Pension Plans)

We use the data in 401K.RAW to estimate the effect of a plan's match rate (*mrate*) on the participation rate (*prate*) in its 401(k) pension plan. The match rate is the amount the firm contributes to a worker's fund for each dollar the worker contributes (up to some limit); thus, *mrate* = .75 means that the firm contributes 75 cents for each dollar contributed by the worker. The participation rate is the percentage of eligible workers having a 401(k) account. The variable *age* is the age of the 401(k) plan. There are 1,534 plans in the data set, the average *prate* is 87.36, the average *mrate* is .732, and the average *age* is 13.2.

Regressing *prate* on *mrate*, *age* gives

$$\widehat{prate} = 80.12 + 5.52\, mrate + .243\, age. \tag{3.23}$$

Thus, both *mrate* and *age* have the expected effects. What happens if we do not control for *age*? The estimated effect of *age* is not trivial, and so we might expect a large change in the estimated effect of *mrate* if *age* is dropped from the regression. However, the simple regression of *prate* on *mrate* yields $\widehat{prate} = 83.08 + 5.86\, mrate$. The simple regression estimate of the effect of *mrate* on *prate* is clearly different from the multiple regression estimate, but the difference is not very big. (The simple regression estimate is only about 6.2 percent larger than the multiple regression estimate.) This can be explained by the fact that the sample correlation between *mrate* and *age* is only .12.

In the case with k independent variables, the simple regression of y on x_1 and the multiple regression of y on $x_1, x_2, \ldots, x_k$ produce an identical estimate of x_1 only if (1) the OLS coefficients on x_2 through x_k are all zero *or* (2) x_1 is uncorrelated with *each of* $x_2, \ldots, x_k$. Neither of these is very likely in practice. But if the coefficients on x_2 through x_k are small, or the sample correlations between x_1 and the other independent variables are insubstantial, then the simple and multiple regression estimates of the effect of x_1 on y can be similar.

Goodness-of-Fit

As with simple regression, we can define the **total sum of squares (SST)**, the **explained sum of squares (SSE)**, and the **residual sum of squares** or **sum of squared residuals (SSR)**, as

$$\text{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.24}$$

$$\text{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.25}$$

$$\text{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \tag{3.26}$$

Using the same argument as in the simple regression case, we can show that

$$\text{SST} = \text{SSE} + \text{SSR}. \tag{3.27}$$

In other words, the total variation in $\{y_i\}$ is the sum of the total variations in $\{\hat{y}_i\}$ and in $\{\hat{u}_i\}$.

Assuming that the total variation in y is nonzero, as is the case unless y_i is constant in the sample, we can divide (3.27) by SST to get

$$\text{SSR/SST} + \text{SSE/SST} = 1.$$

Just as in the simple regression case, the R-squared is defined to be

$$R^2 \equiv \text{SSE/SST} = 1 - \text{SSR/SST}, \tag{3.28}$$

and it is interpreted as the proportion of the sample variation in y_i that is explained by the OLS regression line. By definition, R^2 is a number between zero and one.

R^2 can also be shown to equal the squared correlation coefficient between the actual y_i and the fitted values $\hat{y}_i$. That is,

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2\right)} \tag{3.29}$$

(We have put the average of the $\hat{y}_i$ in (3.29) to be true to the formula for a correlation coefficient; we know that this average equals $\bar{y}$ because the sample average of the residuals is zero and $y_i = \hat{y}_i + \hat{u}_i$.)

An important fact about R^2 is that it never decreases, and it usually increases when another independent variable is added to a regression. This algebraic fact follows because, by definition, the sum of squared residuals never increases when additional regressors are added to the model.

The fact that R^2 never decreases when *any* variable is added to a regression makes it a poor tool for deciding whether one variable or several variables should be added to a model. The factor that should determine whether an explanatory variable belongs in a model is whether the explanatory variable has a nonzero partial effect on y in the *population*. We will show how to test this hypothesis in Chapter 4 when we cover statistical inference. We will also see that, when used properly, R^2 allows us to *test* a group of variables to see if it is important for explaining y. For now, we use it as a goodness-of-fit measure for a given model.

EXAMPLE 3.4
(Determinants of College GPA)

From the grade point average regression that we did earlier, the equation with R^2 is

$$\widehat{colGPA} = 1.29 + .453\, hsGPA + .0094\, ACT$$
$$n = 141,\ R^2 = .176.$$

This means that *hsGPA* and *ACT* together explain about 17.6 percent of the variation in college GPA for this sample of students. This may not seem like a high percentage, but we must remember that there are many other factors—including family background, personality, quality of high school education, affinity for college—that contribute to a student's college performance. If *hsGPA* and *ACT* explained almost all of the variation in *colGPA*, then performance in college would be preordained by high school performance!

EXAMPLE 3.5
(Explaining Arrest Records)

CRIME1.RAW contains data on arrests during the year 1986 and other information on 2,725 men born in either 1960 or 1961 in California. Each man in the sample was arrested at least once prior to 1986. The variable *narr86* is the number of times the man was arrested during 1986, it is zero for most men in the sample (72.29 percent), and it varies from 0 to 12. (The percentage of the men arrested once during 1986 was 20.51.) The variable *pcnv* is the proportion (not percentage) of arrests prior to 1986 that led to conviction, *avgsen* is average sentence length served for prior convictions (zero for most people), *ptime86* is months spent in prison in 1986, and *qemp86* is the number of quarters during which the man was employed in 1986 (from zero to four).

A linear model explaining arrests is

$$narr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 ptime86 + \beta_4 qemp86 + u,$$

where *pcnv* is a proxy for the likelihood for being convicted of a crime and *avgsen* is a measure of expected severity of punishment, if convicted. The variable *ptime86* captures the incarcerative effects of crime: if an individual is in prison, he cannot be arrested for a crime outside of prison. Labor market opportunities are crudely captured by *qemp86*.

First, we estimate the model without the variable *avgsen*. We obtain

$$\widehat{narr86} = .712 - .150\, pcnv - .034\, ptime86 - .104\, qemp86$$
$$n = 2{,}725,\ R^2 = .0413$$

This equation says that, as a group, the three variables *pcnv*, *ptime86*, and *qemp86* explain about 4.1 percent of the variation in *narr86*.

Each of the OLS slope coefficients has the anticipated sign. An increase in the proportion of convictions lowers the predicted number of arrests. If we increase *pcnv* by .50 (a large increase in the probability of conviction), then, holding the other factors fixed, $\Delta\widehat{narr86} = -.150(.5) = -.075$. This may seem unusual because an arrest cannot change by a fraction. But we can use this value to obtain the predicted change in expected arrests for a large group of men. For example, among 100 men, the predicted fall in arrests when *pcnv* increases by .5 is -7.5.

Similarly, a longer prison term leads to a lower predicted number of arrests. In fact, if *ptime86* increases from 0 to 12, predicted arrests for a particular man falls by .034(12) = .408. Another quarter in which legal employment is reported lowers predicted arrests by .104, which would be 10.4 arrests among 100 men.

If *avgsen* is added to the model, we know that R^2 will increase. The estimated equation is

$$\widehat{narr86} = .707 - .151\ pcnv + .0074\ avgsen - .037\ ptime86 - .103\ qemp86$$

$$n = 2{,}725,\ R^2 = .0422.$$

Thus, adding the average sentence variable increases R^2 from .0413 to .0422, a practically small effect. The sign of the coefficient on *avgsen* is also unexpected: it says that a longer average sentence length increases criminal activity.

Example 3.5 deserves a final word of caution. The fact that the four explanatory variables included in the second regression explain only about 4.2 percent of the variation in *narr86* does not necessarily mean that the equation is useless. Even though these variables collectively do not explain much of the variation in arrests, it is still possible that the OLS estimates are reliable estimates of the ceteris paribus effects of each independent variable on *narr86*. As we will see, whether this is the case does not directly depend on the size of R^2. Generally, a low R^2 indicates that it is hard to predict individual outcomes on y with much accuracy, something we study in more detail in Chapter 6. In the arrest example, the small R^2 reflects what we already suspect in the social sciences: it is generally very difficult to predict individual behavior.

Regression Through the Origin

Sometimes, an economic theory or common sense suggests that β_0 should be zero, and so we should briefly mention OLS estimation when the intercept is zero. Specifically, we now seek an equation of the form

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_k x_k, \qquad \textbf{(3.30)}$$

where the symbol "~" over the estimates is used to distinguish them from the OLS estimates obtained along with the intercept [as in (3.11)]. In (3.30), when $x_1 = 0$, $x_2 = 0$, …, $x_k = 0$, the predicted value is zero. In this case, $\tilde{\beta}_1, \ldots, \tilde{\beta}_k$ are said to be the OLS estimates from the regression of y on $x_1, x_2, \ldots, x_k$ *through the origin*.

The OLS estimates in (3.30), as always, minimize the sum of squared residuals, but with the intercept set at zero. You should be warned that the properties of OLS that we derived earlier no longer hold for regression through the origin. In particular, the OLS residuals no longer have a zero sample average. Further, if R^2 is defined as 1 − SSR/SST, where SST is given in (3.24) and SSR is now $\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik})^2$, then R^2 can actually be negative. This means that the sample average, $\bar{y}$, "explains" more of the variation in the y_i than the explanatory variables. Either we should include an intercept in the regression or conclude that the explanatory variables poorly explain y. In order to always have a nonnegative R-squared, some economists prefer to calculate R^2 as the squared correlation coefficient between the actual and fit-

ted values of y, as in (3.29). (In this case, the average fitted value must be computed directly since it no longer equals $\bar{y}$.) However, there is no set rule on computing R-squared for regression through the origin.

One serious drawback with regression through the origin is that, if the intercept β_0 in the population model is different from zero, then the OLS estimators of the slope parameters will be biased. The bias can be severe in some cases. The cost of estimating an intercept when β_0 is truly zero is that the variances of the OLS slope estimators are larger.

3.3 THE EXPECTED VALUE OF THE OLS ESTIMATORS

We now turn to the statistical properties of OLS for estimating the parameters in an underlying population model. In this section, we derive the expected value of the OLS estimators. In particular, we state and discuss four assumptions, which are direct extensions of the simple regression model assumptions, under which the OLS estimators are unbiased for the population parameters. We also explicitly obtain the bias in OLS when an important variable has been omitted from the regression.

You should remember that statistical properties have nothing to do with a particular sample, but rather with the property of estimators when random sampling is done repeatedly. Thus, Sections 3.3, 3.4, and 3.5 are somewhat abstract. While we give examples of deriving bias for particular models, it is not meaningful to talk about the statistical properties of a set of estimates obtained from a single sample.

The first assumption we make simply defines the multiple linear regression (MLR) model.

ASSUMPTION MLR.1 (LINEAR IN PARAMETERS)
The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \tag{3.31}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters (constants) of interest, and u is an unobservable random error or random disturbance term.

Equation (3.31) formally states the **population model**, sometimes called the **true model**, to allow for the possibility that we might estimate a model that differs from (3.31). The key feature is that the model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$. As we know, (3.31) is quite flexible because y and the independent variables can be arbitrary functions of the underlying variables of interest, such as natural logarithms and squares [see, for example, equation (3.7)].

ASSUMPTION MLR.2 (RANDOM SAMPLING)
We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i): i = 1, 2, \ldots, n\}$, from the population model described by (3.31).

Sometimes we need to write the equation for a particular observation i: for a randomly drawn observation from the population, we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i. \quad \textbf{(3.32)}$$

Remember that i refers to the observation, and the second subscript on x is the variable number. For example, we can write a CEO salary equation for a particular CEO i as

$$\log(salary_i) = \beta_0 + \beta_1 \log(sales_i) + \beta_2 ceoten_i + \beta_3 ceoten_i^2 + u_i. \quad \textbf{(3.33)}$$

The term u_i contains the unobserved factors for CEO i that affect his or her salary. For applications, it is usually easiest to write the model in population form, as in (3.31). It contains less clutter and emphasizes the fact that we are interested in estimating a population relationship.

In light of model (3.31), the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ from the regression of y on $x_1, \ldots, x_k$ are now considered to be estimators of $\beta_0, \beta_1, \ldots, \beta_k$. We saw, in Section 3.2, that OLS chooses the estimates for a particular sample so that the residuals average out to zero and the sample correlation between each independent variable and the residuals is zero. For OLS to be unbiased, we need the *population* version of this condition to be true.

ASSUMPTION MLR.3 (ZERO CONDITIONAL MEAN)

The error u has an expected value of zero, given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0. \quad \textbf{(3.34)}$$

One way that Assumption MLR.3 can fail is if the functional relationship between the explained and explanatory variables is misspecified in equation (3.31): for example, if we forget to include the quadratic term inc^2 in the consumption function $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ when we estimate the model. Another functional form misspecification occurs when we use the level of a variable when the log of the variable is what actually shows up in the population model, or vice versa. For example, if the true model has log(*wage*) as the dependent variable but we use *wage* as the dependent variable in our regression analysis, then the estimators will be biased. Intuitively, this should be pretty clear. We will discuss ways of detecting functional form misspecification in Chapter 9.

Omitting an important factor that is correlated with any of $x_1, x_2, \ldots, x_k$ causes Assumption MLR.3 to fail also. With multiple regression analysis, we are able to include many factors among the explanatory variables, and omitted variables are less likely to be a problem in multiple regression analysis than in simple regression analysis. Nevertheless, in any application there are always factors that, due to data limitations or ignorance, we will not be able to include. If we think these factors should be controlled for and they are correlated with one or more of the independent variables, then Assumption MLR.3 will be violated. We will derive this bias in some simple models later.

There are other ways that u can be correlated with an explanatory variable. In Chapter 15, we will discuss the problem of measurement error in an explanatory variable. In Chapter 16, we cover the conceptually more difficult problem in which one or more of the explanatory variables is determined jointly with y. We must postpone our study of these problems until we have a firm grasp of multiple regression analysis under an ideal set of assumptions.

When Assumption MLR.3 holds, we often say we have **exogenous explanatory variables**. If x_j is correlated with u for any reason, then x_j is said to be an **endogenous explanatory variable**. The terms "exogenous" and "endogenous" originated in simultaneous equations analysis (see Chapter 16), but the term "endogenous explanatory variable" has evolved to cover any case where an explanatory variable may be correlated with the error term.

The final assumption we need to show that OLS is unbiased ensures that the OLS estimators are actually well-defined. For simple regression, we needed to assume that the single independent variable was not constant in the sample. The corresponding assumption for multiple regression analysis is more complicated.

ASSUMPTION MLR.4 (NO PERFECT COLLINEARITY)
In the sample (and therefore in the population), none of the independent variables is constant, and there are no *exact linear* relationships among the independent variables.

The no perfect collinearity assumption concerns only the independent variables. Beginning students of econometrics tend to confuse Assumptions MLR.4 and MLR.3, so we emphasize here that MLR.4 says *nothing* about the relationship between u and the explanatory variables.

Assumption MLR.4 is more complicated than its counterpart for simple regression because we must now look at relationships between all independent variables. If an independent variable in (3.31) is an exact linear combination of the other independent variables, then we say the model suffers from **perfect collinearity**, and it cannot be estimated by OLS.

It is important to note that Assumption MLR.4 *does* allow the independent variables to be correlated; they just cannot be *perfectly* correlated. If we did not allow for any correlation among the independent variables, then multiple regression would not be very useful for econometric analysis. For example, in the model relating test scores to educational expenditures and average family income,

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u,$$

we fully expect *expend* and *avginc* to be correlated: school districts with high average family incomes tend to spend more per student on education. In fact, the primary motivation for including *avginc* in the equation is that we suspect it is correlated with *expend*, and so we would like to hold it fixed in the analysis. Assumption MLR.4 only rules out *perfect* correlation between *expend* and *avginc* in our sample. We would be very unlucky to obtain a sample where per student expenditures are perfectly correlated with average family income. But some correlation, perhaps a substantial amount, is expected and certainly allowed.

The simplest way that two independent variables can be perfectly correlated is when one variable is a constant multiple of another. This can happen when a researcher inadvertently puts the same variable measured in different units into a regression equation. For example, in estimating a relationship between consumption and income, it makes no sense to include as independent variables income measured in dollars as well as income measured in thousands of dollars. One of these is redundant. What sense would it make to hold income measured in dollars fixed while changing income measured in thousands of dollars?

We already know that different nonlinear functions of the same variable *can* appear among the regressors. For example, the model $cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$ does not violate Assumption MLR.4: even though $x_2 = inc^2$ is an exact function of $x_1 = inc$, inc^2 is not an exact *linear* function of *inc*. Including inc^2 in the model is a useful way to generalize functional form, unlike including income measured in dollars and in thousands of dollars.

Common sense tells us not to include the same explanatory variable measured in different units in the same regression equation. There are also more subtle ways that one independent variable can be a multiple of another. Suppose we would like to estimate an extension of a constant elasticity consumption function. It might seem natural to specify a model such as

$$\log(cons) = \beta_0 + \beta_1 \log(inc) + \beta_2 \log(inc^2) + u, \tag{3.35}$$

where $x_1 = \log(inc)$ and $x_2 = \log(inc^2)$. Using the basic properties of the natural log (see Appendix A), $\log(inc^2) = 2 \cdot \log(inc)$. That is, $x_2 = 2x_1$, and naturally this holds for all observations in the sample. This violates Assumption MLR.4. What we should do instead is include $[\log(inc)]^2$, not $\log(inc^2)$, along with $\log(inc)$. This is a sensible extension of the constant elasticity model, and we will see how to interpret such models in Chapter 6.

Another way that independent variables can be perfectly collinear is when one independent variable can be expressed as an exact linear function of two or more of the other independent variables. For example, suppose we want to estimate the effect of campaign spending on campaign outcomes. For simplicity, assume that each election has two candidates. Let *voteA* be the percent of the vote for Candidate A, let *expendA* be campaign expenditures by Candidate A, let *expendB* be campaign expenditures by Candidate B, and let *totexpend* be total campaign expenditures; the latter three variables are all measured in dollars. It may seem natural to specify the model as

$$voteA = \beta_0 + \beta_1 expendA + \beta_2 expendB + \beta_3 totexpend + u, \tag{3.36}$$

in order to isolate the effects of spending by each candidate and the total amount of spending. But this model violates Assumption MLR.4 because $x_3 = x_1 + x_2$ by definition. Trying to interpret this equation in a ceteris paribus fashion reveals the problem. The parameter of β_1 in equation (3.36) is supposed to measure the effect of increasing expenditures by Candidate A by one dollar on Candidate A's vote, holding Candidate B's spending *and* total spending fixed. This is nonsense, because if *expendB* and *totexpend* are held fixed, then we cannot increase *expendA*.

The solution to the perfect collinearity in (3.36) is simple: drop any one of the three variables from the model. We would probably drop *totexpend*, and then the coefficient on *expendA* would measure the effect of increasing expenditures by A on the percentage of the vote received by A, holding the spending by B fixed.

The prior examples show that Assumption MLR.4 can fail if we are not careful in specifying our model. Assumption MLR.4 also fails if the sample size, n, is too small in relation to the number of parameters being estimated. In the general regression model in equation (3.31), there are $k + 1$ parameters, and MLR.4 fails if $n < k + 1$. Intuitively, this makes sense: to estimate $k + 1$ parameters, we need at least $k + 1$ observations. Not surprisingly, it is better to have as many observations as possible, something we will see with our variance calculations in Section 3.4.

QUESTION 3.3

In the previous example, if we use as explanatory variables *expendA*, *expendB*, and *shareA*, where *shareA* = 100·(*expendA*/*totexpend*) is the percentage share of total campaign expenditures made by Candidate A, does this violate Assumption MLR.4?

If the model is carefully specified and $n \geq k + 1$, Assumption MLR.4 can fail in rare cases due to bad luck in collecting the sample. For example, in a wage equation with education and experience as variables, it is possible that we could obtain a random sample where each individual has exactly twice as much education as years of experience. This scenario would cause Assumption MLR.4 to fail, but it can be considered very unlikely unless we have an extremely small sample size.

We are now ready to show that, under these four multiple regression assumptions, the OLS estimators are unbiased. As in the simple regression case, the expectations are conditional on the values of the independent variables in the sample, but we do not show this conditioning explicitly.

THEOREM 3.1 (UNBIASEDNESS OF OLS)
Under Assumptions MLR.1 through MLR.4,

$$\mathrm{E}(\hat{\beta}_j) = \beta_j, j = 0,1, \ldots, k, \qquad \textbf{(3.37)}$$

for any values of the population parameter β_j. In other words, the OLS estimators are unbiased estimators of the population parameters.

In our previous empirical examples, Assumption MLR.4 has been satisfied (since we have been able to compute the OLS estimates). Furthermore, for the most part, the samples are randomly chosen from a well-defined population. If we believe that the specified models are correct under the key Assumption MLR.3, then we can conclude that OLS is unbiased in these examples.

Since we are approaching the point where we can use multiple regression in serious empirical work, it is useful to remember the meaning of unbiasedness. It is tempting, in examples such as the wage equation in equation (3.19), to say something like "9.2 percent is an unbiased estimate of the return to education." As we know, an estimate cannot be unbiased: an estimate is a fixed number, obtained from a particular sample, which usually is not equal to the population parameter. When we say that OLS is unbi-

ased under Assumptions MLR.1 through MLR.4, we mean that the *procedure* by which the OLS estimates are obtained is unbiased when we view the procedure as being applied across all possible random samples. We hope that we have obtained a sample that gives us an estimate close to the population value, but, unfortunately, this cannot be assured.

Including Irrelevant Variables in a Regression Model

One issue that we can dispense with fairly quicky is that of **inclusion of an irrelevant variable** or **overspecifying the model** in multiple regression analysis. This means that one (or more) of the independent variables is included in the model even though it has no partial effect on y in the population. (That is, its population coefficient is zero.)

To illustrate the issue, suppose we specify the model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \tag{3.38}$$

and this model satisfies Assumptions MLR.1 through MLR.4. However, x_3 has no effect on y after x_1 and x_2 have been controlled for, which means that $\beta_3 = 0$. The variable x_3 may or may not be correlated with x_1 or x_2; all that matters is that, once x_1 and x_2 are controlled for, x_3 has no effect on y. In terms of conditional expectations, $E(y|x_1,x_2,x_3) = E(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Because we do not know that $\beta_3 = 0$, we are inclined to estimate the equation including x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3. \tag{3.39}$$

We have included the irrelevant variable, x_3, in our regression. What is the effect of including x_3 in (3.39) when its coefficient in the population model (3.38) is zero? In terms of the unbiasedness of $\hat{\beta}_1$ and $\hat{\beta}_2$, there is *no effect*. This conclusion requires no special derivation, as it follows immediately from Theorem 3.1. Remember, unbiasedness means $E(\hat{\beta}_j) = \beta_j$ for *any* value of β_j, including $\beta_j = 0$. Thus, we can conclude that $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$, and $E(\hat{\beta}_3) = 0$ (for any values of β_0, β_1, and β_2). Even though $\hat{\beta}_3$ itself will never be exactly zero, its average value across many random samples will be zero.

The conclusion of the preceding example is much more general: including one or more irrelevant variables in a multiple regression model, or overspecifying the model, does not affect the unbiasedness of the OLS estimators. Does this mean it is harmless to include irrelevant variables? No. As we will see in Section 3.4, including irrelevant variables can have undesirable effects on the *variances* of the OLS estimators.

Omitted Variable Bias: The Simple Case

Now suppose that, rather than including an irrelevant variable, we omit a variable that actually belongs in the true (or population) model. This is often called the problem of **excluding a relevant variable** or **underspecifying the model**. We claimed in Chapter 2 and earlier in this chapter that this problem generally causes the OLS estimators to be biased. It is time to show this explicitly and, just as importantly, to derive the direction and size of the bias.

Deriving the bias caused by omitting an important variable is an example of **misspecification analysis**. We begin with the case where the true population model has two explanatory variables and an error term:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.40}$$

and we assume that this model satisfies Assumptions MLR.1 through MLR.4.

Suppose that our primary interest is in β_1, the partial effect of x_1 on y. For example, y is hourly wage (or log of hourly wage), x_1 is education, and x_2 is a measure of innate ability. In order to get an unbiased estimator of β_1, we *should* run a regression of y on x_1 and x_2 (which gives unbiased estimators of β_0, β_1, and β_2). However, due to our ignorance or data inavailability, we estimate the model by *excluding* x_2. In other words, we perform a simple regression of y on x_1 only, obtaining the equation

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.41}$$

We use the symbol "~" rather than "^" to emphasize that $\tilde{\beta}_1$ comes from an underspecified model.

When first learning about the omitted variables problem, it can be difficult for the student to distinguish between the underlying true model, (3.40) in this case, and the model that we actually estimate, which is captured by the regression in (3.41). It may seem silly to omit the variable x_2 if it belongs in the model, but often we have no choice. For example, suppose that *wage* is determined by

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u. \tag{3.42}$$

Since ability is not observed, we instead estimate the model

$$wage = \beta_0 + \beta_1 educ + v,$$

where $v = \beta_2 abil + u$. The estimator of β_1 from the simple regression of *wage* on *educ* is what we are calling $\tilde{\beta}_1$.

We derive the expected value of $\tilde{\beta}_1$ conditional on the sample values of x_1 and x_2. Deriving this expectation is not difficult because $\tilde{\beta}_1$ is just the OLS slope estimator from a simple regression, and we have already studied this estimator extensively in Chapter 2. The difference here is that we must analyze its properties when the simple regression model is misspecified due to an omitted variable.

From equation (2.49), we can express $\tilde{\beta}_1$ as

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) y_i}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}. \tag{3.43}$$

The next step is the most important one. Since (3.40) is the true model, we write y for each observation i as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \tag{3.44}$$

(*not* $y_i = \beta_0 + \beta_1 x_{i1} + u_i$, because the true model contains x_2). Let SST_1 be the denominator in (3.43). If we plug (3.44) in for y_i in (3.43), the numerator in (3.43) becomes

$$\begin{aligned}
&\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) \\
&= \beta_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2} + \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) u_i \\
&\equiv \beta_1 \text{SST}_1 + \beta_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2} + \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) u_i. \qquad (3.45)
\end{aligned}$$

If we divide (3.45) by SST_1, take the expectation conditional on the values of the independent variables, and use $\text{E}(u_i) = 0$, we obtain

$$\text{E}(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}. \tag{3.46}$$

Thus, $\text{E}(\tilde{\beta}_1)$ does not generally equal β_1: $\tilde{\beta}_1$ is biased for β_1.

The ratio multiplying β_2 in (3.46) has a simple interpretation: it is just the slope coefficient from the regression of x_2 on x_1, using our sample on the independent variables, which we can write as

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1. \tag{3.47}$$

Because we are conditioning on the sample values of both independent variables, $\tilde{\delta}_1$ is not random here. Therefore, we can write (3.46) as

$$\text{E}(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1, \tag{3.48}$$

which implies that the bias in $\tilde{\beta}_1$ is $\text{E}(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$. This is often called the **omitted variable bias**.

From equation (3.48), we see that there are two cases where $\tilde{\beta}_1$ is unbiased. The first is pretty obvious: if $\beta_2 = 0$—so that x_2 does not appear in the true model (3.40)—then $\tilde{\beta}_1$ is unbiased. We already know this from the simple regression analysis in Chapter 2. The second case is more interesting. If $\tilde{\delta}_1 = 0$, then $\tilde{\beta}_1$ is unbiased for β_1, even if $\beta_2 \neq 0$.

Since $\tilde{\delta}_1$ is the sample covariance between x_1 and x_2 over the sample variance of x_1, $\tilde{\delta}_1 = 0$ if, and only if, x_1 and x_2 are uncorrelated in the sample. Thus, we have the important conclusion that, if x_1 and x_2 are uncorrelated in the sample, then $\tilde{\beta}_1$ is unbiased. This is not surprising: in Section 3.2, we showed that the simple regression estimator $\tilde{\beta}_1$ and the multiple regression estimator $\hat{\beta}_1$ are the same when x_1 and x_2 are uncorrelated in the sample. [We can also show that $\tilde{\beta}_1$ is unbiased without conditioning on the x_{i2} if

Table 3.2

Summary of Bias in $\tilde{\beta}_1$ When x_2 is Omitted in Estimating Equation (3.40)

	Corr(x_1,x_2) > 0	**Corr(x_1,x_2) < 0**
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

$E(x_2|x_1) = E(x_2)$; then, for estimating β_1, leaving x_2 in the error term does not violate the zero conditional mean assumption for the error, once we adjust the intercept.]

When x_1 and x_2 are correlated, $\tilde{\delta}_1$ has the same sign as the correlation between x_1 and x_2: $\tilde{\delta}_1 > 0$ if x_1 and x_2 are positively correlated and $\tilde{\delta}_1 < 0$ if x_1 and x_2 are negatively correlated. The sign of the bias in $\tilde{\beta}_1$ depends on the signs of both β_2 and $\tilde{\delta}_1$ and is summarized in Table 3.2 for the four possible cases when there is bias. Table 3.2 warrants careful study. For example, the bias in $\tilde{\beta}_1$ is positive if $\beta_2 > 0$ (x_2 has a positive effect on y) and x_1 and x_2 are positively correlated. The bias is negative if $\beta_2 > 0$ and x_1 and x_2 are negatively correlated. And so on.

Table 3.2 summarizes the direction of the bias, but the size of the bias is also very important. A small bias of either sign need not be a cause for concern. For example, if the return to education in the population is 8.6 percent and the bias in the OLS estimator is 0.1 percent (a tenth of one percentage point), then we would not be very concerned. On the other hand, a bias on the order of three percentage points would be much more serious. The size of the bias is determined by the sizes of β_2 and $\tilde{\delta}_1$.

In practice, since β_2 is an unknown population parameter, we cannot be certain whether β_2 is positive or negative. Nevertheless, we usually have a pretty good idea about the direction of the partial effect of x_2 on y. Further, even though the sign of the correlation between x_1 and x_2 cannot be known if x_2 is not observed, in many cases we can make an educated guess about whether x_1 and x_2 are positively or negatively correlated.

In the wage equation (3.42), by definition more ability leads to higher productivity and therefore higher wages: $\beta_2 > 0$. Also, there are reasons to believe that *educ* and *abil* are positively correlated: on average, individuals with more innate ability choose higher levels of education. Thus, the OLS estimates from the simple regression equation $wage = \beta_0 + \beta_1 educ + v$ are *on average* too large. This does not mean that the estimate obtained from our sample is too big. We can only say that if we collect many random samples and obtain the simple regression estimates each time, then the average of these estimates will be greater than β_1.

EXAMPLE 3.6
(Hourly Wage Equation)

Suppose the model $\log(wage) = \beta_0 + \beta_1 educ + \beta_2 abil + u$ satisfies Assumptions MLR.1 through MLR.4. The data set in WAGE1.RAW does not contain data on ability, so we estimate β_1 from the simple regression

$$\widehat{\log(wage)} = .584 + .083\ educ$$
$$n = 526,\ R^2 = .186.$$

This is only the result from a single sample, so we cannot say that .083 is greater than β_1; the true return to education could be lower or higher than 8.3 percent (and we will never know for sure). Nevertheless, we know that the average of the estimates across all random samples would be too large.

As a second example, suppose that, at the elementary school level, the average score for students on a standardized exam is determined by

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 povrate + u,$$

where *expend* is expenditure per student and *povrate* is the poverty rate of the children in the school. Using school district data, we only have observations on the percent of students with a passing grade and per student expenditures; we do not have information on poverty rates. Thus, we estimate β_1 from the simple regression of *avgscore* on *expend*.

We can again obtain the likely bias in $\tilde{\beta}_1$. First, β_2 is probably negative: there is ample evidence that children living in poverty score lower, on average, on standardized tests. Second, the average expenditure per student is probably negatively correlated with the poverty rate: the higher the poverty rate, the lower the average per-student spending, so that $\text{Corr}(x_1,x_2) < 0$. From Table 3.2, $\tilde{\beta}_1$ will have a positive bias. This observation has important implications. It could be that the true effect of spending is zero; that is, $\beta_1 = 0$. However, the simple regression estimate of β_1 will usually be greater than zero, and this could lead us to conclude that expenditures are important when they are not.

When reading and performing empirical work in economics, it is important to master the terminology associated with biased estimators. In the context of omitting a variable from model (3.40), if $\text{E}(\tilde{\beta}_1) > \beta_1$, then we say that $\tilde{\beta}_1$ has an **upward bias**. When $\text{E}(\tilde{\beta}_1) < \beta_1$, $\tilde{\beta}_1$ has a **downward bias**. These definitions are the same whether β_1 is positive or negative. The phrase **biased towards zero** refers to cases where $\text{E}(\tilde{\beta}_1)$ is closer to zero than β_1. Therefore, if β_1 is positive, then $\tilde{\beta}_1$ is biased towards zero if it has a downward bias. On the other hand, if $\beta_1 < 0$, then $\tilde{\beta}_1$ is biased towards zero if it has an upward bias.

Omitted Variable Bias: More General Cases

Deriving the sign of omitted variable bias when there are multiple regressors in the estimated model is more difficult. We must remember that correlation between a single explanatory variable and the error generally results in *all* OLS estimators being biased. For example, suppose the population model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \qquad \textbf{(3.49)}$$

satisfies Assumptions MLR.1 through MLR.4. But we omit x_3 and estimate the model as

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2. \quad \textbf{(3.50)}$$

Now, suppose that x_2 and x_3 are uncorrelated, but that x_1 is correlated with x_3. In other words, x_1 is correlated with the omitted variable, but x_2 is not. It is tempting to think that, while $\tilde{\beta}_1$ is probably biased based on the derivation in the previous subsection, $\tilde{\beta}_2$ is unbiased because x_2 is uncorrelated with x_3. Unfortunately, this is *not* generally the case: both $\tilde{\beta}_1$ and $\tilde{\beta}_2$ will normally be biased. The only exception to this is when x_1 and x_2 are also uncorrelated.

Even in the fairly simple model above, it is difficult to obtain the direction of the bias in $\tilde{\beta}_1$ and $\tilde{\beta}_2$. This is because x_1, x_2, and x_3 can all be pairwise correlated. Nevertheless, an approximation is often practically useful. If we assume that x_1 and x_2 are uncorrelated, then we can study the bias in $\tilde{\beta}_1$ as if x_2 were absent from both the population and the estimated models. In fact, when x_1 and x_2 are uncorrelated, it can be shown that

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i3}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}.$$

This is just like equation (3.46), but β_3 replaces β_2 and x_3 replaces x_2. Therefore, the bias in $\tilde{\beta}_1$ is obtained by replacing β_2 with β_3 and x_2 with x_3 in Table 3.2. If $\beta_3 > 0$ and $\mathrm{Corr}(x_1,x_3) > 0$, the bias in $\tilde{\beta}_1$ is positive. And so on.

As an example, suppose we add *exper* to the wage model:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u.$$

If *abil* is omitted from the model, the estimators of both β_1 and β_2 are biased, even if we assume *exper* is uncorrelated with *abil*. We are mostly interested in the return to education, so it would be nice if we could conclude that $\tilde{\beta}_1$ has an upward or downward bias due to omitted ability. This conclusion is not possible without further assumptions. As an *approximation*, let us suppose that, in addition to *exper* and *abil* being uncorrelated, *educ* and *exper* are also uncorrelated. (In reality, they are somewhat negatively correlated.) Since $\beta_3 > 0$ and *educ* and *abil* are positively correlated, $\tilde{\beta}_1$ would have an upward bias, just as if *exper* were not in the model.

The reasoning used in the previous example is often followed as a rough guide for obtaining the likely bias in estimators in more complicated models. Usually, the focus is on the relationship between a particular explanatory variable, say x_1, and the key omitted factor. Strictly speaking, ignoring all other explanatory variables is a valid practice only when each one is uncorrelated with x_1, but it is still a useful guide.

3.4 THE VARIANCE OF THE OLS ESTIMATORS

We now obtain the variance of the OLS estimators so that, in addition to knowing the central tendencies of $\hat{\beta}_j$, we also have a measure of the spread in its sampling distribution. Before finding the variances, we add a homoskedasticity assumption, as in Chapter 2. We do this for two reasons. First, the formulas are simplified by imposing the con-

stant error variance assumption. Second, in Section 3.5, we will see that OLS has an important efficiency property if we add the homoskedasticity assumption.

In the multiple regression framework, homoskedasticity is stated as follows:

ASSUMPTION MLR.5 (HOMOSKEDASTICITY)
$\text{Var}(u|x_1,\ldots,x_k) = \sigma^2$.

Assumption MLR.5 means that the variance in the error term, u, conditional on the explanatory variables, is the *same* for all combinations of outcomes of the explanatory variables. If this assumption fails, then the model exhibits heteroskedasticity, just as in the two-variable case.

In the equation

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u,$$

homoskedasticity requires that the variance of the unobserved error u does not depend on the levels of education, experience, or tenure. That is,

$$\text{Var}(u|educ, exper, tenure) = \sigma^2.$$

If this variance changes with any of the three explanatory variables, then heteroskedasticity is present.

Assumptions MLR.1 through MLR.5 are collectively known as the **Gauss-Markov assumptions** (for cross-sectional regression). So far, our statements of the assumptions are suitable only when applied to cross-sectional analysis with random sampling. As we will see, the Gauss-Markov assumptions for time series analysis, and for other situations such as panel data analysis, are more difficult to state, although there are many similarities.

In the discussion that follows, we will use the symbol $\boldsymbol{x}$ to denote the set of all independent variables, $(x_1, \ldots, x_k)$. Thus, in the wage regression with *educ*, *exper*, and *tenure* as independent variables, $\boldsymbol{x}$ = (*educ*, *exper*, *tenure*). Now we can write Assumption MLR.3 as

$$\text{E}(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

and Assumption MLR.5 is the same as $\text{Var}(y|\boldsymbol{x}) = \sigma^2$. Stating the two assumptions in this way clearly illustrates how Assumption MLR.5 differs greatly from Assumption MLR.3. Assumption MLR.3 says that the expected value of y, given $\boldsymbol{x}$, is linear in the parameters, but it certainly depends on $x_1, x_2, \ldots, x_k$. Assumption MLR.5 says that the variance of y, given $\boldsymbol{x}$, does *not* depend on the values of the independent variables.

We can now obtain the variances of the $\hat{\beta}_j$, where we again condition on the sample values of the independent variables. The proof is in the appendix to this chapter.

THEOREM 3.2 (SAMPLING VARIANCES OF THE OLS SLOPE ESTIMATORS)
Under Assumptions MLR.1 through MLR.5, conditional on the sample values of the independent variables,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}, \tag{3.51}$$

for $j = 1,2,\ldots,k$, where $\text{SST}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j, and R_j^2 is the R-squared from regressing x_j on all other independent variables (and including an intercept).

Before we study equation (3.51) in more detail, it is important to know that all of the Gauss-Markov assumptions are used in obtaining this formula. While we did not need the homoskedasticity assumption to conclude that OLS is unbiased, we do need it to validate equation (3.51).

The size of $\text{Var}(\hat{\beta}_j)$ is practically important. A larger variance means a less precise estimator, and this translates into larger confidence intervals and less accurate hypotheses tests (as we will see in Chapter 4). In the next subsection, we discuss the elements comprising (3.51).

The Components of the OLS Variances: Multicollinearity

Equation (3.51) shows that the variance of $\hat{\beta}_j$ depends on three factors: σ^2, SST_j, and R_j^2. Remember that the index j simply denotes any one of the independent variables (such as education or poverty rate). We now consider each of the factors affecting $\text{Var}(\hat{\beta}_j)$ in turn.

THE ERROR VARIANCE, σ^2. From equation (3.51), a larger σ^2 means larger variances for the OLS estimators. This is not at all surprising: more "noise" in the equation (a larger σ^2) makes it more difficult to estimate the partial effect of any of the independent variables on y, and this is reflected in higher variances for the OLS slope estimators. Since σ^2 is a feature of the population, it has nothing to do with the sample size. It is the one component of (3.51) that is unknown. We will see later how to obtain an unbiased estimator of σ^2.

For a given dependent variable y, there is really only one way to reduce the error variance, and that is to add more explanatory variables to the equation (take some factors out of the error term). This is not always possible, nor is it always desirable for reasons discussed later in the chapter.

THE TOTAL SAMPLE VARIATION IN x_j, SST_j. From equation (3.51), the larger the total variation in x_j, the smaller is $\text{Var}(\hat{\beta}_j)$. Thus, everything else being equal, for estimating β_j we prefer to have as much sample variation in x_j as possible. We already discovered this in the simple regression case in Chapter 2. While it is rarely possible for us to choose the sample values of the independent variables, there *is* a way to increase the sample variation in each of the independent variables: increase the sample size. In fact, when sampling randomly from a population, SST_j increases without bound as the sample size gets larger and larger. This is the component of the variance that systematically depends on the sample size.

When SST_j is small, $Var(\hat{\beta}_j)$ can get very large, but a small SST_j is not a violation of Assumption MLR.4. Technically, as SST_j goes to zero, $Var(\hat{\beta}_j)$ approaches infinity. The extreme case of no sample variation in x_j, $SST_j = 0$, is not allowed by Assumption MLR.4.

THE LINEAR RELATIONSHIPS AMONG THE INDEPENDENT VARIABLES, R_j^2. The term R_j^2 in equation (3.51) is the most difficult of the three components to understand. This term does not appear in simple regression analysis because there is only one independent variable in such cases. It is important to see that this R-squared is distinct from the R-squared in the regression of y on $x_1, x_2, \ldots, x_k$: R_j^2 is obtained from a regression involving only the independent variables in the original model, where x_j plays the role of a dependent variable.

Consider first the $k = 2$ case: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Then $Var(\hat{\beta}_1) = \sigma^2/[SST_1(1 - R_1^2)]$, where R_1^2 is the R-squared from the simple regression of x_1 on x_2 (and an intercept, as always). Since the R-squared measures goodness-of-fit, a value of R_1^2 close to one indicates that x_2 explains much of the variation in x_1 in the sample. This means that x_1 and x_2 are highly correlated.

As R_1^2 increases to one, $Var(\hat{\beta}_1)$ gets larger and larger. Thus, a high degree of linear relationship between x_1 and x_2 can lead to large variances for the OLS slope estimators. (A similar argument applies to $\hat{\beta}_2$.) See Figure 3.1 for the relationship between $Var(\hat{\beta}_1)$ and the R-squared from the regression of x_1 on x_2.

In the general case, R_j^2 is the proportion of the total variation in x_j that can be explained by the *other* independent variables appearing in the equation. For a given σ^2 and SST_j, the smallest $Var(\hat{\beta}_j)$ is obtained when $R_j^2 = 0$, which happens if, and only if, x_j has zero sample correlation with *every other* independent variable. This is the best case for estimating β_j, but it is rarely encountered.

The other extreme case, $R_j^2 = 1$, is ruled out by Assumption MLR.4, because $R_j^2 = 1$ means that, in the sample, x_j is a *perfect* linear combination of some of the other independent variables in the regression. A more relevant case is when R_j^2 is "close" to one. From equation (3.51) and Figure 3.1, we see that this can cause $Var(\hat{\beta}_j)$ to be large: $Var(\hat{\beta}_j) \to \infty$ as $R_j^2 \to 1$. High (but not perfect) correlation between two or more of the independent variables is called **multicollinearity**.

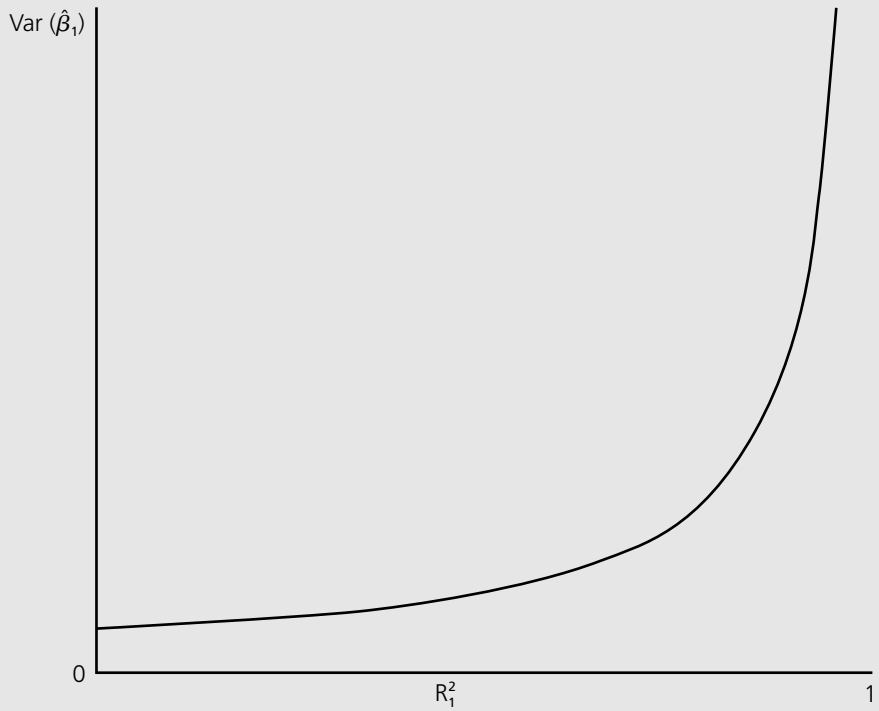
Before we discuss the multicollinearity issue further, it is important to be very clear on one thing: a case where R_j^2 is close to one is *not* a violation of Assumption MLR.4.

Since multicollinearity violates none of our assumptions, the "problem" of multicollinearity is not really well-defined. When we say that multicollinearity arises for estimating β_j when R_j^2 is "close" to one, we put "close" in quotation marks because there is no absolute number that we can cite to conclude that multicollinearity is a problem. For example, $R_j^2 = .9$ means that 90 percent of the sample variation in x_j can be explained by the other independent variables in the regression model. Unquestionably, this means that x_j has a strong linear relationship to the other independent variables. But whether this translates into a $Var(\hat{\beta}_j)$ that is too large to be useful depends on the sizes of σ^2 and SST_j. As we will see in Chapter 4, for statistical inference, what ultimately matters is how big $\hat{\beta}_j$ is in relation to its standard deviation.

Just as a large value of R_j^2 can cause large $Var(\hat{\beta}_j)$, so can a small value of SST_j. Therefore, a small sample size can lead to large sampling variances, too. Worrying

Figure 3.1

Var $(\hat{\beta}_1)$ as a function of R_1^2.



about high degrees of correlation among the independent variables in the sample is really no different from worrying about a small sample size: both work to increase Var$(\hat{\beta}_j)$. The famous University of Wisconsin econometrician Arthur Goldberger, reacting to econometricians' obsession with multicollinearity, has [tongue-in-cheek] coined the term **micronumerosity**, which he defines as the "problem of small sample size." [For an engaging discussion of multicollinearity and micronumerosity, see Goldberger (1991).]

Although the problem of multicollinearity cannot be clearly defined, one thing is clear: everything else being equal, for estimating β_j it is better to have less correlation between x_j and the other independent variables. This observation often leads to a discussion of how to "solve" the multicollinearity problem. In the social sciences, where we are usually passive collectors of data, there is no good way to reduce variances of unbiased estimators other than to collect more data. For a given data set, we can try dropping other independent variables from the model in an effort to reduce multicollinearity. Unfortunately, dropping a variable that belongs in the population model can lead to bias, as we saw in Section 3.3.

Perhaps an example at this point will help clarify some of the issues raised concerning multicollinearity. Suppose we are interested in estimating the effect of various

school expenditure categories on student performance. It is likely that expenditures on teacher salaries, instructional materials, athletics, and so on, are highly correlated: wealthier schools tend to spend more on everything, and poorer schools spend less on everything. Not surprisingly, it can be difficult to estimate the effect of any particular expenditure category on student performance when there is little variation in one category that cannot largely be explained by variations in the other expenditure categories (this leads to high R_j^2 for each of the expenditure variables). Such multicollinearity problems can be mitigated by collecting more data, but in a sense we have imposed the problem on ourselves: we are asking questions that may be too subtle for the available data to answer with any precision. We can probably do much better by changing the scope of the analysis and lumping all expenditure categories together, since we would no longer be trying to estimate the partial effect of each separate category.

Another important point is that a high degree of correlation between certain independent variables can be irrelevant as to how well we can estimate other parameters in the model. For example, consider a model with three independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

where x_2 and x_3 are highly correlated. Then $\text{Var}(\hat{\beta}_2)$ and $\text{Var}(\hat{\beta}_3)$ may be large. But the amount of correlation between x_2 and x_3 has no direct effect on $\text{Var}(\hat{\beta}_1)$. In fact, if x_1 is uncorrelated with x_2 and x_3, then $R_1^2 = 0$ and $\text{Var}(\hat{\beta}_1) = \sigma^2/\text{SST}_1$, regardless of how much correlation there is between x_2 and x_3. If β_1 is the parameter of interest, we do not really care about the amount of correlation between x_2 and x_3.

QUESTION 3.4

Suppose you postulate a model explaining final exam score in terms of class attendance. Thus, the dependent variable is final exam score, and the key explanatory variable is number of classes attended. To control for student abilities and efforts outside the classroom, you include among the explanatory variables cumulative GPA, SAT score, and measures of high school performance. Someone says, "You cannot hope to learn anything from this exercise because cumulative GPA, SAT score, and high school performance are likely to be highly collinear." What should be your response?

The previous observation is important because economists often include many controls in order to isolate the causal effect of a particular variable. For example, in looking at the relationship between loan approval rates and percent of minorities in a neighborhood, we might include variables like average income, average housing value, measures of creditworthiness, and so on, because these factors need to be accounted for in order to draw causal conclusions about discrimination. Income, housing prices, and creditworthiness are generally highly correlated with each other. But high correlations among these variables do not make it more difficult to determine the effects of discrimination.

Variances in Misspecified Models

The choice of whether or not to include a particular variable in a regression model can be made by analyzing the tradeoff between bias and variance. In Section 3.3, we derived the bias induced by leaving out a relevant variable when the true model contains two explanatory variables. We continue the analysis of this model by comparing the variances of the OLS estimators.

Write the true population model, which satisfies the Gauss-Markov assumptions, as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

We consider two estimators of β_1. The estimator $\hat{\beta}_1$ comes from the multiple regression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \tag{3.52}$$

In other words, we include x_2, along with x_1, in the regression model. The estimator $\tilde{\beta}_1$ is obtained by omitting x_2 from the model and running a simple regression of y on x_1:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.53}$$

When $\beta_2 \neq 0$, equation (3.53) excludes a relevant variable from the model and, as we saw in Section 3.3, this induces a bias in $\tilde{\beta}_1$ unless x_1 and x_2 are uncorrelated. On the other hand, $\hat{\beta}_1$ is unbiased for β_1 for any value of β_2, including $\beta_2 = 0$. It follows that, if bias is used as the only criterion, $\hat{\beta}_1$ is preferred to $\tilde{\beta}_1$.

The conclusion that $\hat{\beta}_1$ is always preferred to $\tilde{\beta}_1$ does not carry over when we bring variance into the picture. Conditioning on the values of x_1 and x_2 in the sample, we have, from (3.51),

$$\text{Var}(\hat{\beta}_1) = \sigma^2/[\text{SST}_1(1 - R_1^2)], \tag{3.54}$$

where SST_1 is the total variation in x_1, and R_1^2 is the R-squared from the regression of x_1 on x_2. Further, a simple modification of the proof in Chapter 2 for two-variable regression shows that

$$\text{Var}(\tilde{\beta}_1) = \sigma^2/\text{SST}_1. \tag{3.55}$$

Comparing (3.55) to (3.54) shows that $\text{Var}(\tilde{\beta}_1)$ is always *smaller* than $\text{Var}(\hat{\beta}_1)$, unless x_1 and x_2 are uncorrelated in the sample, in which case the two estimators $\tilde{\beta}_1$ and $\hat{\beta}_1$ are the same. Assuming that x_1 and x_2 are not uncorrelated, we can draw the following conclusions:

1. When $\beta_2 \neq 0$, $\tilde{\beta}_1$ is biased, $\hat{\beta}_1$ is unbiased, and $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.
2. When $\beta_2 = 0$, $\tilde{\beta}_1$ and $\hat{\beta}_1$ are both unbiased, and $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.

From the second conclusion, it is clear that $\tilde{\beta}_1$ is preferred if $\beta_2 = 0$. Intuitively, if x_2 does not have a partial effect on y, then including it in the model can only exacerbate the multicollinearity problem, which leads to a less efficient estimator of β_1. A higher variance for the estimator of β_1 is the cost of including an irrelevant variable in a model.

The case where $\beta_2 \neq 0$ is more difficult. Leaving x_2 out of the model results in a biased estimator of β_1. Traditionally, econometricians have suggested comparing the likely size of the bias due to omitting x_2 with the reduction in the variance—summarized in the size of R_1^2—to decide whether x_2 should be included. However, when $\beta_2 \neq 0$, there are two favorable reasons for including x_2 in the model. The most important of these is that any bias in $\tilde{\beta}_1$ does not shrink as the sample size grows; in fact, the bias does not necessarily follow any pattern. Therefore, we can usefully think of the bias as being roughly the same for any sample size. On the other hand, $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$ both shrink to zero as n gets large, which means that the multicollinearity induced by adding x_2 becomes less important as the sample size grows. In large samples, we would prefer $\hat{\beta}_1$.

The other reason for favoring $\hat{\beta}_1$ is more subtle. The variance formula in (3.55) is conditional on the values of x_{i1} and x_{i2} in the sample, which provides the best scenario for $\tilde{\beta}_1$. When $\beta_2 \neq 0$, the variance of $\tilde{\beta}_1$ conditional only on x_1 is larger than that presented in (3.55). Intuitively, when $\beta_2 \neq 0$ and x_2 is excluded from the model, the error variance increases because the error effectively contains part of x_2. But formula (3.55) ignores the error variance increase because it treats both regressors as nonrandom. A full discussion of which independent variables to condition on would lead us too far astray. It is sufficient to say that (3.55) is too generous when it comes to measuring the precision in $\tilde{\beta}_1$.

Estimating σ^2: Standard Errors of the OLS Estimators

We now show how to choose an unbiased estimator of σ^2, which then allows us to obtain unbiased estimators of $\text{Var}(\hat{\beta}_j)$.

Since $\sigma^2 = \text{E}(u^2)$, an unbiased "estimator" of σ^2 is the sample average of the squared errors: $n^{-1}\sum_{i=1}^{n} u_i^2$. Unfortunately, this is not a true estimator because we do not observe the u_i. Nevertheless, recall that the errors can be written as $u_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}$, and so the reason we do not observe the u_i is that we do not know the β_j. When we replace each β_j with its OLS estimator, we get the OLS residuals:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}.$$

It seems natural to estimate σ^2 by replacing u_i with the $\hat{u}_i$. In the simple regression case, we saw that this leads to a biased estimator. The unbiased estimator of σ^2 in the general multiple regression case is

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} \hat{u}_i^2\right) \Big/ (n - k - 1) \equiv \text{SSR}/(n - k - 1). \qquad \textbf{(3.56)}$$

We already encountered this estimator in the $k = 1$ case in simple regression.

The term $n - k - 1$ in (3.56) is the **degrees of freedom** (***df***) for the general OLS problem with n observations and k independent variables. Since there are $k + 1$ parameters in a regression model with k independent variables and an intercept, we can write

$$\begin{aligned} df &= n - (k + 1) \\ &= (\text{number of observations}) - (\text{number of estimated parameters}). \end{aligned} \qquad \textbf{(3.57)}$$

This is the easiest way to compute the degrees of freedom in a particular application: count the number of parameters, including the intercept, and subtract this amount from the number of observations. (In the rare case that an intercept is not estimated, the number of parameters decreases by one.)

Technically, the division by $n - k - 1$ in (3.56) comes from the fact that the expected value of the sum of squared residuals is $\text{E}(\text{SSR}) = (n - k - 1)\sigma^2$. Intuitively, we can figure out why the degrees of freedom adjustment is necessary by returning to the first order conditions for the OLS estimators. These can be written as $\sum_{i=1}^{n} \hat{u}_i = 0$ and

$\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, where $j = 1,2, \ldots, k$. Thus, in obtaining the OLS estimates, $k + 1$ restrictions are imposed on the OLS residuals. This means that, given $n - (k + 1)$ of the residuals, the remaining $k + 1$ residuals are known: there are only $n - (k + 1)$ degrees of freedom in the residuals. (This can be contrasted with the *errors* u_i, which have n degrees of freedom in the sample.)

For reference, we summarize this discussion with Theorem 3.3. We proved this theorem for the case of simple regression analysis in Chapter 2 (see Theorem 2.3). (A general proof that requires matrix algebra is provided in Appendix E.)

THEOREM 3.3 (UNBIASED ESTIMATION OF σ^2)
Under the Gauss-Markov Assumptions MLR.1 through MLR.5, $E(\hat{\sigma}^2) = \sigma^2$.

The positive square root of $\hat{\sigma}^2$, denoted $\hat{\sigma}$, is called the **standard error of the regression** or **SER**. The SER is an estimator of the standard deviation of the error term. This estimate is usually reported by regression packages, although it is called different things by different packages. (In addition to ser, $\hat{\sigma}$ is also called the *standard error of the estimate* and the *root mean squared error*.)

Note that $\hat{\sigma}$ can either decrease or increase when another independent variable is added to a regression (for a given sample). This is because, while SSR must fall when another explanatory variable is added, the degrees of freedom also falls by one. Because SSR is in the numerator and *df* is in the denominator, we cannot tell beforehand which effect will dominate.

For constructing confidence intervals and conducting tests in Chapter 4, we need to estimate the **standard deviation of $\hat{\beta}_j$**, which is just the square root of the variance:

$$\text{sd}(\hat{\beta}_j) = \sigma/[\text{SST}_j(1 - R_j^2)]^{1/2}.$$

Since σ is unknown, we replace it with its estimator, $\hat{\sigma}$. This gives us the **standard error of $\hat{\beta}_j$**:

$$\text{se}(\hat{\beta}_j) = \hat{\sigma}/[\text{SST}_j(1 - R_j^2)]^{1/2}. \tag{3.58}$$

Just as the OLS estimates can be obtained for any given sample, so can the standard errors. Since $\text{se}(\hat{\beta}_j)$ depends on $\hat{\sigma}$, the standard error has a sampling distribution, which will play a role in Chapter 4.

We should emphasize one thing about standard errors. Because (3.58) is obtained directly from the variance formula in (3.51), and because (3.51) relies on the homoskedasticity Assumption MLR.5, it follows that the standard error formula in (3.58) is *not* a valid estimator of $\text{sd}(\hat{\beta}_j)$ if the errors exhibit heteroskedasticity. Thus, while the presence of heteroskedasticity does not cause bias in the $\hat{\beta}_j$, it does lead to bias in the usual formula for $\text{Var}(\hat{\beta}_j)$, which then invalidates the standard errors. This is important because any regression package computes (3.58) as the default standard error for each coefficient (with a somewhat different representation for the intercept). If we suspect heteroskedasticity, then the "usual" OLS standard errors are invalid and some corrective action should be taken. We will see in Chapter 8 what methods are available for dealing with heteroskedasticity.

3.5 EFFICIENCY OF OLS: THE GAUSS-MARKOV THEOREM

In this section, we state and discuss the important **Gauss-Markov Theorem**, which justifies the use of the OLS method rather than using a variety of competing estimators. We know one justification for OLS already: under Assumptions MLR.1 through MLR.4, OLS is unbiased. However, there are *many* unbiased estimators of the β_j under these assumptions (for example, see Problem 3.12). Might there be other unbiased estimators with variances smaller than the OLS estimators?

If we limit the class of competing estimators appropriately, then we can show that OLS *is* best within this class. Specifically, we will argue that, under Assumptions MLR.1 through MLR.5, the OLS estimator $\hat{\beta}_j$ for β_j is the **best linear unbiased estimator** (**BLUE**). In order to state the theorem, we need to understand each component of the acronym "BLUE." First, we know what an estimator is: it is a rule that can be applied to any sample of data to produce an estimate. We also know what an unbiased estimator is: in the current context, an estimator, say $\tilde{\beta}_j$, of β_j is an unbiased estimator of β_j if $\mathrm{E}(\tilde{\beta}_j) = \beta_j$ for any $\beta_0, \beta_1, \ldots, \beta_k$.

What about the meaning of the term "linear"? In the current context, an estimator $\tilde{\beta}_j$ of β_j is linear if, and only if, it can be expressed as a linear function of the data on the dependent variable:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i, \tag{3.59}$$

where each w_{ij} can be a function of the sample values of all the independent variables. The OLS estimators are linear, as can be seen from equation (3.22).

Finally, how do we define "best"? For the current theorem, best is defined as *smallest variance*. Given two unbiased estimators, it is logical to prefer the one with the smallest variance (see Appendix C).

Now, let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ denote the OLS estimators in the model (3.31) under Assumptions MLR.1 through MLR.5. The Gauss-Markov theorem says that, for any estimator $\tilde{\beta}_j$ which is *linear* and *unbiased*, $\mathrm{Var}(\hat{\beta}_j) \leq \mathrm{Var}(\tilde{\beta}_j)$, and the inequality is usually strict. In other words, in the class of linear unbiased estimators, OLS has the smallest variance (under the five Gauss-Markov assumptions). Actually, the theorem says more than this. If we want to estimate any linear function of the β_j, then the corresponding linear combination of the OLS estimators achieves the smallest variance among all linear unbiased estimators. We conclude with a theorem, which is proven in Appendix 3A.

THEOREM 3.4 (GAUSS-MARKOV THEOREM)
Under Assumptions MLR.1 through MLR.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the best linear unbiased estimators (BLUEs) of $\beta_0, \beta_1, \ldots, \beta_k$, respectively.

It is because of this theorem that Assumptions MLR.1 through MLR.5 are known as the Gauss-Markov assumptions (for cross-sectional analysis).

The importance of the Gauss-Markov theorem is that, when the standard set of assumptions holds, we need not look for alternative unbiased estimators of the form (3.59): none will be better than OLS. Equivalently, if we are presented with an estimator that is both linear and unbiased, then we know that the variance of this estimator is at least as large as the OLS variance; no additional calculation is needed to show this.

For our purposes, Theorem 3.4 justifies the use of OLS to estimate multiple regression models. If any of the Gauss-Markov assumptions fail, then this theorem no longer holds. We already know that failure of the zero conditional mean assumption (Assumption MLR.3) causes OLS to be biased, so Theorem 3.4 also fails. We also know that heteroskedasticity (failure of Assumption MLR.5) does not cause OLS to be biased. However, OLS no longer has the smallest variance among linear unbiased estimators in the presence of heteroskedasticity. In Chapter 8, we analyze an estimator that improves upon OLS when we know the brand of heteroskedasticity.

SUMMARY

1. The multiple regression model allows us to effectively hold other factors fixed while examining the effects of a particular independent variable on the dependent variable. It explicitly allows the independent variables to be correlated.

2. Although the model is linear in its *parameters*, it can be used to model nonlinear relationships by appropriately choosing the dependent and independent variables.

3. The method of ordinary least squares is easily applied to the multiple regression model. Each slope estimate measures the partial effect of the corresponding independent variable on the dependent variable, holding all other independent variables fixed.

4. R^2 is the proportion of the sample variation in the dependent variable explained by the independent variables, and it serves as a goodness-of-fit measure. It is important not to put too much weight on the value of R^2 when evaluating econometric models.

5. Under the first four Gauss-Markov assumptions (MLR.1 through MLR.4), the OLS estimators are unbiased. This implies that including an irrelevant variable in a model has no effect on the unbiasedness of the intercept and other slope estimators. On the other hand, omitting a relevant variable causes OLS to be biased. In many circumstances, the direction of the bias can be determined.

6. Under the five Gauss-Markov assumptions, the variance of an OLS slope estimator is given by $\text{Var}(\hat{\beta}_j) = \sigma^2/[\text{SST}_j(1 - R_j^2)]$. As the error variance σ^2 increases, so does $\text{Var}(\hat{\beta}_j)$, while $\text{Var}(\hat{\beta}_j)$ decreases as the sample variation in x_j, SST_j, increases. The term R_j^2 measures the amount of collinearity between x_j and the other explanatory variables. As R_j^2 approaches one, $\text{Var}(\hat{\beta}_j)$ is unbounded.

7. Adding an irrelevant variable to an equation generally increases the variances of the remaining OLS estimators because of multicollinearity.

8. Under the Gauss-Markov assumptions (MLR.1 through MLR.5), the OLS estimators are best linear unbiased estimators (BLUE).

KEY TERMS

Best Linear Unbiased Estimator (BLUE)
Biased Towards Zero
Ceteris Paribus
Degrees of Freedom (*df*)
Disturbance
Downward Bias
Endogenous Explanatory Variable
Error Term
Excluding a Relevant Variable
Exogenous Explanatory Variables
Explained Sum of Squares (SSE)
First Order Conditions
Gauss-Markov Assumptions
Gauss-Markov Theorem
Inclusion of an Irrelevant Variable
Intercept
Micronumerosity
Misspecification Analysis
Multicollinearity
Multiple Linear Regression Model
Multiple Regression Analysis
Omitted Variable Bias
OLS Intercept Estimate
OLS Regression Line
OLS Slope Estimate
Ordinary Least Squares
Overspecifying the Model
Partial Effect
Perfect Collinearity
Population Model
Residual
Residual Sum of Squares
Sample Regression Function (SRF)
Slope Parameters
Standard Deviation of $\hat{\beta}_j$
Standard Error of $\hat{\beta}_j$
Standard Error of the Regression (SER)
Sum of Squared Residuals (SSR)
Total Sum of Squares (SST)
True Model
Underspecifying the Model
Upward Bias

PROBLEMS

3.1 Using the data in GPA2.RAW on 4,137 college students, the following equation was estimated by OLS:

$$\widehat{colgpa} = 1.392 - .0135\ hsperc + .00148\ sat$$

$$n = 4{,}137,\ R^2 = .273,$$

where *colgpa* is measured on a four-point scale, *hsperc* is the percentile in the high school graduating class (defined so that, for example, *hsperc* = 5 means the *top* five percent of the class), and *sat* is the combined math and verbal scores on the student achievement test.

(i) Why does it make sense for the coefficient on *hsperc* to be negative?

(ii) What is the predicted college GPA when *hsperc* = 20 and *sat* = 1050?

(iii) Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?

(iv) Holding *hsperc* fixed, what difference in SAT scores leads to a predicted *colgpa* difference of .50, or one-half of a grade point? Comment on your answer.

3.2 The data in WAGE2.RAW on working men was used to estimate the following equation:

$$\widehat{educ} = 10.36 - .094\, sibs + .131\, meduc + .210\, feduc$$

$$n = 722,\ R^2 = .214,$$

where *educ* is years of schooling, *sibs* is number of siblings, *meduc* is mother's years of schooling, and *feduc* is father's years of schooling.

(i) Does *sibs* have the expected effect? Explain. Holding *meduc* and *feduc* fixed, by how much does *sibs* have to increase to reduce predicted years of education by one year? (A noninteger answer is acceptable here.)

(ii) Discuss the interpretation of the coefficient on *meduc*.

(iii) Suppose that Man A has no siblings, and his mother and father each have 12 years of education. Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

3.3 The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years. (See also Problem 2.12.)

(i) If adults trade off sleep for work, what is the sign of β_1?

(ii) What signs do you think β_2 and β_3 will have?

(iii) Using the data in SLEEP75.RAW, the estimated equation is

$$\widehat{sleep} = 3638.25 - .148\, totwrk - 11.13\, educ + 2.20\, age$$

$$n = 706,\ R^2 = .113.$$

If someone works five more hours per week, by how many minutes is *sleep* predicted to fall? Is this a large tradeoff?

(iv) Discuss the sign and magnitude of the estimated coefficient on *educ*.

(v) Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?

3.4 The median starting salary for new law school graduates is determined by

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

where *LSAT* is median LSAT score for the graduating class, *GPA* is the median college GPA for the class, *libvol* is the number of volumes in the law school library, *cost* is the annual cost of attending law school, and *rank* is a law school ranking (with *rank* = 1 being the best).

(i) Explain why we expect $\beta_5 \leq 0$.

(ii) What signs to you expect for the other slope parameters? Justify your answers.

(iii) Using the data in LAWSCH85.RAW, the estimated equation is

$$\log(\hat{salary}) = 8.34 + .0047\ LSAT + .248\ GPA + .095 \log(libvol)$$
$$+ .038 \log(cost) - .0033\ rank$$
$$n = 136,\ R^2 = .842.$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (Report your answer as a percent.)

(iv) Interpret the coefficient on the variable log(*libvol*).

(v) Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?

3.5 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student the sum of hours in the four activities must be 168.

(i) In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

(ii) Explain why this model violates Assumption MLR.4.

(iii) How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.4?

3.6 Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You are interested in estimating the sum of the parameters on x_1 and x_2; call this $\theta_1 = \beta_1 + \beta_2$. Show that $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ is an unbiased estimator of θ_1.

3.7 Which of the following can cause OLS estimators to be biased?

(i) Heteroskedasticity.

(ii) Omitting an important variable.

(iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

3.8 Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*):

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u.$$

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that *avgtrain* and *avgabil* are negatively correlated, what is the likely bias in $\tilde{\beta}_1$ obtained from the simple regression of *avgprod* on *avgtrain*?

3.9 The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (*rooms*):

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 rooms + u.$$

(i) What are the probable signs of β_1 and β_2? What is the interpretation of β_1? Explain.

(ii) Why might *nox* [more precisely, log(*nox*)] and *rooms* be negatively correlated? If this is the case, does the simple regression of log(*price*) on log(*nox*) produce an upward or downward biased estimator of β_1?

(iii) Using the data in HPRICE2.RAW, the following equations were estimated:

$$\widehat{\log(price)} = 11.71 - 1.043 \log(nox), \; n = 506, \; R^2 = .264.$$

$$\widehat{\log(price)} = 9.23 - .718 \log(nox) + .306 \, rooms, \; n = 506, \; R^2 = .514.$$

Is the relationship between the simple and multiple regression estimates of the elasticity of *price* with respect to *nox* what you would have predicted, given your answer in part (ii)? Does this mean that $-.718$ is definitely closer to the true elasticity than -1.043?

3.10 Suppose that the population model determining y is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

and this model satisifies the Gauss-Markov assumptions. However, we estimate the model that omits x_3. Let $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$ be the OLS estimators from the regression of y on x_1 and x_2. Show that the expected value of $\tilde{\beta}_1$ (given the values of the independent variables in the sample) is

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

where the $\hat{r}_{i1}$ are the OLS residuals from the regression of x_1 on x_2. [*Hint:* The formula for $\tilde{\beta}_1$ comes from equation (3.22). Plug $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ into this equation. After some algebra, take the expectation treating x_{i3} and $\hat{r}_{i1}$ as nonrandom.]

3.11 The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of counties in the United States:

$$growth = \beta_0 + \beta_1 share_{\mathrm{P}} + \beta_2 share_{\mathrm{I}} + \beta_3 share_{\mathrm{S}} + other\ factors,$$

where *growth* is the percentage change in employment from 1980 to 1990, $share_{\mathrm{P}}$ is the share of property taxes in total tax revenue, $share_{\mathrm{I}}$ is the share of income tax revenues,

and $share_S$ is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, $share_F$, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

(i) Why must we omit one of the tax share variables from the equation?

(ii) Give a careful interpretation of β_1.

3.12 (i) Consider the simple regression model $y = \beta_0 + \beta_1 x + u$ under the first four Gauss-Markov assumptions. For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log(1 + x^2)$, define $z_i = g(x_i)$. Define a slope estimator as

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} (z_i - \bar{z})y_i\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z})x_i\right).$$

Show that $\tilde{\beta}_1$ is linear and unbiased. Remember, because $E(u|x) = 0$, you can treat both x_i and z_i as nonrandom in your derivation.

(ii) Add the homoskedasticity assumption, MLR.5. Show that

$$\text{Var}(\tilde{\beta}_1) = \sigma^2 \left(\sum_{i=1}^{n} (z_i - \bar{z})^2\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z})x_i\right)^2.$$

(iii) Show directly that, under the Gauss-Markov assumptions, $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. [Hint: The Cauchy-Schwartz inequality in Appendix B implies that

$$\left(n^{-1}\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})\right)^2 \leq \left(n^{-1}\sum_{i=1}^{n} (z_i - \bar{z})^2\right)\left(n^{-1}\sum_{i=1}^{n} (x_i - \bar{x})^2\right);$$

notice that we can drop $\bar{x}$ from the sample covariance.]

COMPUTER EXERCISES

3.13 A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth rate that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 faminc + u.$$

(i) What is the most likely sign for β_2?

(ii) Do you think *cigs* and *faminc* are likely to be correlated? Explain why the correlation might be positive or negative.

(iii) Now estimate the equation with and without *faminc*, using the data in BWGHT.RAW. Report the results in equation form, including the sample size and *R*-squared. Discuss your results, focusing on whether

adding *faminc* substantially changes the estimated effect of *cigs* on *bwght*.

3.14 Use the data in HPRICE1.RAW to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

(v) The first house in the sample has $sqrft = 2{,}438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was \$300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

3.15 The file CEOSAL2.RAW contains data on 177 chief executive officers, which can be used to examine the effects of firm performance on CEO salary.

(i) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.

(ii) Add *profits* to the model from part (i). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?

(iii) Add the variable *ceoten* to the model in part (ii). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?

(iv) Find the sample correlation coefficient between the variables log(*mktval*) and *profits*. Are these variables highly correlated? What does this say about the OLS estimators?

3.16 Use the data in ATTEND.RAW for this exercise.

(i) Obtain the minimum, maximum, and average values for the variables *atndrte*, *priGPA*, and *ACT*.

(ii) Estimate the model

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u$$

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

(iii) Discuss the estimated slope coefficients. Are there any surprises?

(iv) What is the predicted *atndrte*, if $priGPA = 3.65$ and $ACT = 20$? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

(v) If Student A has *priGPA* = 3.1 and *ACT* = 21 and Student B has *priGPA* = 2.1 and *ACT* = 26, what is the predicted difference in their attendance rates?

3.17 Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialling out for Example 3.2. This first requires regressing *educ* on *exper* and *tenure*, and saving the residuals, $\hat{r}_1$. Then, regress log(*wage*) on $\hat{r}_1$. Compare the coefficient on $\hat{r}_1$ with the coefficient on *educ* in the regression of log(*wage*) on *educ*, *exper*, and *tenure*.

APPENDIX 3A

3A.1 Derivation of the First Order Conditions, Equations (3.13)

The analysis is very similar to the simple regression case. We must characterize the solutions to the problem

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2.$$

Taking the partial derivatives with respect to each of the b_j (see Appendix A), evaluating them at the solutions, and setting them equal to zero gives

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1\text{v}} - \ldots - \hat{\beta}_k x_{ik}) = 0$$
$$-2 \sum_{i=1}^{n} x_{ij}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0, \; j = 1, \ldots, k.$$

Cancelling the -2 gives the first order conditions in (3.13).

3A.2 Derivation of Equation (3.22)

To derive (3.22), write x_{i1} in terms of its fitted value and its residual from the regression of x_1 on to $x_2, \ldots, x_k$: $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, $i = 1, \ldots, n$. Now, plug this into the second equation in (3.13):

$$\sum_{i=1}^{n} (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0. \qquad \textbf{(3.60)}$$

By the definition of the OLS residual $\hat{u}_i$, since $\hat{x}_{i1}$ is just a linear function of the explanatory variables $x_{i2}, \ldots, x_{ik}$, it follows that $\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = 0$. Therefore, (3.60) can be expressed as

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0. \qquad \textbf{(3.61)}$$

Since the $\hat{r}_{i1}$ are the residuals from regressing x_1 onto $x_2, \ldots, x_k$, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$ for $j = 2, \ldots, k$. Therefore, (3.61) is equivalent to $\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0$. Finally, we use the fact that $\sum_{i=1}^{n} \hat{x}_{i1}\hat{r}_{i1} = 0$, which means that $\hat{\beta}_1$ solves

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0.$$

Now straightforward algebra gives (3.22), provided, of course, that $\sum_{i=1}^{n} \hat{r}_{i1}^2 > 0$; this is ensured by Assumption MLR.4.

3A.3 Proof of Theorem 3.1

We prove Theorem 3.1 for $\hat{\beta}_1$; the proof for the other slope parameters is virtually identical. (See Appendix E for a more succinct proof using matrices.) Under Assumption MLR.4, the OLS estimators exist, and we can write $\hat{\beta}_1$ as in (3.22). Under Assumption MLR.1, we can write y_i as in (3.32); substitute this for y_i in (3.22). Then, using $\sum_{i=1}^{n} \hat{r}_{i1} = 0$, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$ for all $j = 2, \ldots, k$, and $\sum_{i=1}^{n} x_{i1}\hat{r}_{i1} = \sum_{i=1}^{n} \hat{r}_{i1}^2$, we have

$$\hat{\beta}_1 = \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}u_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \tag{3.62}$$

Now, under Assumptions MLR.2 and MLR.4, the expected value of each u_i, given all independent variables in the sample, is zero. Since the $\hat{r}_{i1}$ are just functions of the sample independent variables, it follows that

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1|\boldsymbol{X}) &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}\mathrm{E}(u_i|\boldsymbol{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) \\ &= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}\cdot 0\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) = \beta_1, \end{aligned}$$

where $\boldsymbol{X}$ denotes the data on all independent variables and $\mathrm{E}(\hat{\beta}_1|\boldsymbol{X})$ is the expected value of $\hat{\beta}_1$, given $x_{i1}, \ldots, x_{ik}$ for all $i = 1, \ldots, n$. This completes the proof.

3A.4 Proof of Theorem 3.2

Again, we prove this for $j = 1$. Write $\hat{\beta}_1$ as in equation (3.62). Now, under MLR.5, $\mathrm{Var}(u_i|\boldsymbol{X}) = \sigma^2$ for all $i = 1, \ldots, n$. Under random sampling, the u_i are independent, even conditional on $\boldsymbol{X}$, and the $\hat{r}_{i1}$ are nonrandom conditional on $\boldsymbol{X}$. Therefore,

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_1|\boldsymbol{X}) &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\,\mathrm{Var}(u_i|\boldsymbol{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)^2 \\ &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\sigma^2\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)^2 = \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \end{aligned}$$

Now, since $\sum_{i=1}^{n} \hat{r}_{i1}^2$ is the sum of squared residuals from regressing x_1 on to $x_2, \ldots, x_k$, $\sum_{i=1}^{n} \hat{r}_{i1}^2 = \text{SST}_1(1 - R_1^2)$. This completes the proof.

3A.5 Proof of Theorem 3.4

We show that, for any other linear unbiased estimator $\tilde{\beta}_1$ of β_1, $\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. The focus on $j = 1$ is without loss of generality.

For $\tilde{\beta}_1$ as in equation (3.59), we can plug in for y_i to obtain

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik} + \sum_{i=1}^{n} w_{i1}u_i.$$

Now, since the w_{i1} are functions of the x_{ij},

$$\begin{aligned}\text{E}(\tilde{\beta}_1|\boldsymbol{X}) &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik} + \sum_{i=1}^{n} w_{i1}\text{E}(u_i|\boldsymbol{X}) \\ &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + \ldots + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik}\end{aligned}$$

because $\text{E}(u_i|\boldsymbol{X}) = 0$, for all $i = 1, \ldots, n$ under MLR.3. Therefore, for $\text{E}(\tilde{\beta}_1|\boldsymbol{X})$ to equal β_1 for any values of the parameters, we must have

$$\sum_{i=1}^{n} w_{i1} = 0, \sum_{i=1}^{n} w_{i1}x_{i1} = 1, \sum_{i=1}^{n} w_{i1}x_{ij} = 0, j = 2, \ldots, k. \tag{3.63}$$

Now, let $\hat{r}_{i1}$ be the residuals from the regression of x_{i1} on to $x_{i2}, \ldots, x_{ik}$. Then, from (3.63), it follows that

$$\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} = 1. \tag{3.64}$$

Now, consider the difference between $\text{Var}(\tilde{\beta}_1|\boldsymbol{X})$ and $\text{Var}(\hat{\beta}_1|\boldsymbol{X})$ under MLR.1 through MLR.5:

$$\sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \tag{3.65}$$

Because of (3.64), we can write the difference in (3.65), without σ^2, as

$$\sum_{i=1}^{n} w_{i1}^2 - \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1} \right)^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \tag{3.66}$$

But (3.66) is simply

$$\sum_{i=1}^{n} (w_{i1} - \hat{\gamma}_1\hat{r}_{i1})^2, \tag{3.67}$$

where $\hat{\gamma}_1 = \left(\sum_{i=1}^{n} w_{i1}\hat{r}_{i1}\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)$, as can be seen by squaring each term in (3.67), summing, and then cancelling terms. Because (3.67) is just the sum of squared residuals from the simple regression of w_{i1} on to $\hat{r}_{i1}$—remember that the sample average of $\hat{r}_{i1}$ is zero—(3.67) must be nonnegative. This completes the proof.

Chapter Four

Multiple Regression Analysis: Inference

This chapter continues our treatment of multiple regression analysis. We now turn to the problem of testing hypotheses about the parameters in the population regression model. We begin by finding the distributions of the OLS estimators under the added assumption that the population error is normally distributed. Sections 4.2 and 4.3 cover hypothesis testing about individual parameters, while Section 4.4 discusses how to test a single hypothesis involving more than one parameter. We focus on testing multiple restrictions in Section 4.5 and pay particular attention to determining whether a group of independent variables can be omitted from a model.

4.1 SAMPLING DISTRIBUTIONS OF THE OLS ESTIMATORS

Up to this point, we have formed a set of assumptions under which OLS is unbiased, and we have also derived and discussed the bias caused by omitted variables. In Section 3.4, we obtained the variances of the OLS estimators under the Gauss-Markov assumptions. In Section 3.5, we showed that this variance is smallest among linear unbiased estimators.

Knowing the expected value and variance of the OLS estimators is useful for describing the precision of the OLS estimators. However, in order to perform statistical inference, we need to know more than just the first two moments of $\hat{\beta}_j$; we need to know the full sampling distribution of the $\hat{\beta}_j$. Even under the Gauss-Markov assumptions, the distribution of $\hat{\beta}_j$ can have virtually any shape.

When we condition on the values of the independent variables in our sample, it is clear that the sampling distributions of the OLS estimators depend on the underlying distribution of the errors. To make the sampling distributions of the $\hat{\beta}_j$ tractable, we now assume that the unobserved error is *normally distributed* in the population. We call this the **normality assumption**.

ASSUMPTION MLR.6 (NORMALITY)

The population error u is *independent* of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0,\sigma^2)$.

Assumption MLR.6 is much stronger than any of our previous assumptions. In fact, since u is independent of the x_j under MLR.6, $E(u|x_1, \ldots, x_k) = E(u) = 0$, and $Var(u|x_1, \ldots, x_k) = Var(u) = \sigma^2$. Thus, if we make Assumption MLR.6, then we are necessarily assuming MLR.3 and MLR.5. To emphasize that we are assuming more than before, we will refer to the the full set of assumptions MLR.1 through MLR.6.

For cross-sectional regression applications, the six assumptions MLR.1 through MLR.6 are called the **classical linear model (CLM) assumptions**. Thus, we will refer to the model under these six assumptions as the **classical linear model**. It is best to think of the CLM assumptions as containing all of the Gauss-Markov assumptions *plus* the assumption of a normally distributed error term.

Under the CLM assumptions, the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ have a stronger efficiency property than they would under the Gauss-Markov assumptions. It can be shown that the OLS estimators are the **minimum variance unbiased estimators**, which means that OLS has the smallest variance among unbiased estimators; we no longer have to restrict our comparison to estimators that are linear in the y_i. This property of OLS under the CLM assumptions is discussed further in Appendix E.

A succinct way to summarize the population assumptions of the CLM is

$$y|\boldsymbol{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2),$$

where $\boldsymbol{x}$ is again shorthand for $(x_1, \ldots, x_k)$. Thus, conditional on $\boldsymbol{x}$, y has a normal distribution with mean linear in $x_1, \ldots, x_k$ and a constant variance. For a single independent variable x, this situation is shown in Figure 4.1.

The argument justifying the normal distribution for the errors usually runs something like this: Because u is the sum of many different unobserved factors affecting y, we can invoke the central limit theorem (see Appendix C) to conclude that u has an approximate normal distribution. This argument has some merit, but it is not without weaknesses. First, the factors in u can have very different distributions in the population (for example, ability and quality of schooling in the error in a wage equation). While the central limit theorem (CLT) can still hold in such cases, the normal approximation can be poor depending on how many factors appear in u and how different are their distributions.

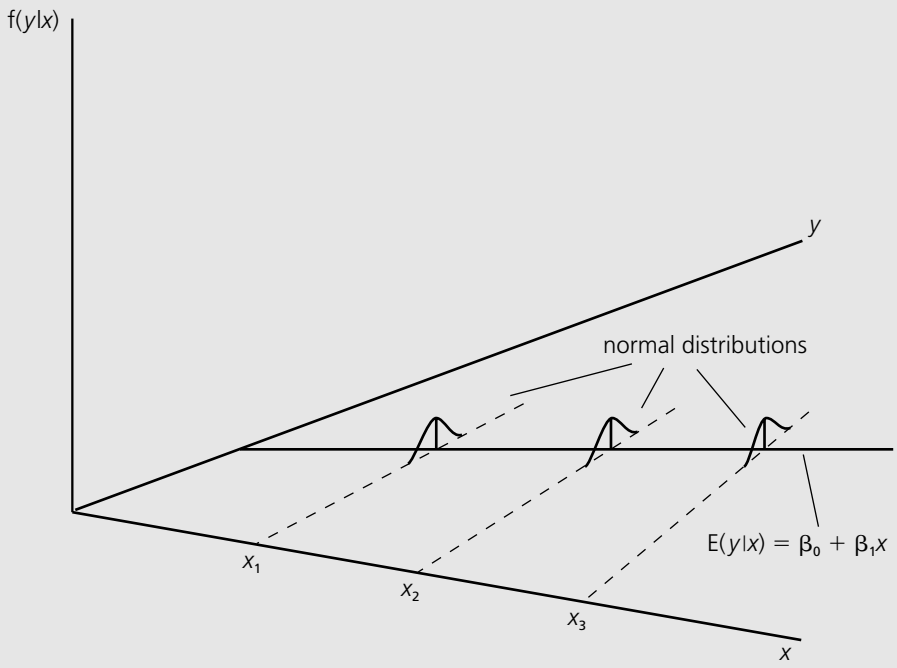
A more serious problem with the CLT argument is that it assumes that all unobserved factors affect y in a separate, additive fashion. Nothing guarantees that this is so. If u is a complicated function of the unobserved factors, then the CLT argument does not really apply.

In any application, whether normality of u can be assumed is really an empirical matter. For example, there is no theorem that says *wage* conditional on *educ*, *exper*, and *tenure* is normally distributed. If anything, simple reasoning suggests that the opposite is true: since *wage* can never be less than zero, it cannot, strictly speaking, have a normal distribution. Further, since there are minimum wage laws, some fraction of the population earns exactly the minimum wage, which also violates the normality assumption. Nevertheless, as a practical matter we can ask whether the conditional wage distribution is "close" to being normal. Past empirical evidence suggests that normality is *not* a good assumption for wages.

Often, using a transformation, especially taking the log, yields a distribution that is closer to normal. For example, something like log(*price*) tends to have a distribution

Figure 4.1

The homoskedastic normal distribution with a single explanatory variable.



that looks more normal than the distribution of *price*. Again, this is an empirical issue, which we will discuss further in Chapter 5.

There are some examples where MLR.6 is clearly false. Whenever y takes on just a few values, it cannot have anything close to a normal distribution. The dependent variable in Example 3.5 provides a good example. The variable *narr86*, the number of times a young man was arrested in 1986, takes on a small range of integer values and is zero for most men. Thus, *narr86* is far from being normally distributed. What can be done in these cases? As we will see in Chapter 5—and this is important—nonnormality of the errors is not a serious problem with large sample sizes. For now, we just make the normality assumption.

Normality of the error term translates into normal sampling distributions of the OLS estimators:

THEOREM 4.1 (NORMAL SAMPLING DISTRIBUTIONS)

Under the CLM assumptions MLR.1 through MLR.6, conditional on the sample values of the independent variables,

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\hat{\beta}_j)], \tag{4.1}$$

where $\text{Var}(\hat{\beta}_j)$ was given in Chapter 3 [equation (3.51)]. Therefore,

$$(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1).$$

The proof of (4.1) is not that difficult, given the properties of normally distributed random variables in Appendix B. Each $\hat{\beta}_j$ can be written as $\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij}u_i$, where $w_{ij} = \hat{r}_{ij}/\text{SSR}_j$, $\hat{r}_{ij}$ is the i^{th} residual from the regression of the x_j on all the other independent variables, and SSR_j is the sum of squared residuals from this regression [see equation (3.62)]. Since the w_{ij} depend only on the independent variables, they can be treated as nonrandom. Thus, $\hat{\beta}_j$ is just a linear combination of the errors in the sample, $\{u_i: i = 1,2, \ldots, n\}$. Under Assumption MLR.6 (and the random sampling Assumption MLR.2), the errors are independent, identically distributed Normal$(0,\sigma^2)$ random variables. An important fact about independent normal random variables is that a linear combination of such random variables is normally distributed (see Appendix B). This basically completes the proof. In Section 3.3, we showed that $\text{E}(\hat{\beta}_j) = \beta_j$, and we derived $\text{Var}(\hat{\beta}_j)$ in Section 3.4; there is no need to re-derive these facts.

QUESTION 4.1

Suppose that u is independent of the explanatory variables, and it takes on the values -2, -1, 0, 1, and 2 with equal probability of 1/5. Does this violate the Gauss-Markov assumptions? Does this violate the CLM assumptions?

The second part of this theorem follows immediately from the fact that when we standardize a normal random variable by dividing it by its standard deviation, we end up with a standard normal random variable.

The conclusions of Theorem 4.1 can be strengthened. In addition to (4.1), any linear combination of the $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ is also normally distributed, and any subset of the $\hat{\beta}_j$ has a *joint* normal distribution. These facts underlie the testing results in the remainder of this chapter. In Chapter 5, we will show that the normality of the OLS estimators is still *approximately* true in large samples even without normality of the errors.

4.2 TESTING HYPOTHESES ABOUT A SINGLE POPULATION PARAMETER: THE *t* TEST

This section covers the very important topic of testing hypotheses about any single parameter in the population regression function. The population model can be written as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad \textbf{(4.2)}$$

and we assume that it satisfies the CLM assumptions. We know that OLS produces unbiased estimators of the β_j. In this section, we study how to test hypotheses about a particular β_j. For a full understanding of hypothesis testing, one must remember that the β_j are unknown features of the population, and we will never know them with certainty. Nevertheless, we can *hypothesize* about the value of β_j and then use statistical inference to test our hypothesis.

In order to construct hypotheses tests, we need the following result:

THEOREM 4.2 (*t* DISTRIBUTION FOR THE STANDARDIZED ESTIMATORS)
Under the CLM assumptions MLR.1 through MLR.6,

$$(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \sim t_{n-k-1}, \tag{4.3}$$

where $k + 1$ is the number of unknown parameters in the population model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ (k slope parameters and the intercept β_0).

This result differs from Theorem 4.1 in some notable respects. Theorem 4.1 showed that, under the CLM assumptions, $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \sim \text{Normal}(0,1)$. The t distribution in (4.3) comes from the fact that the constant σ in $\text{sd}(\hat{\beta}_j)$ has been replaced with the random variable $\hat{\sigma}$. The proof that this leads to a t distribution with $n - k - 1$ degrees of freedom is not especially insightful. Essentially, the proof shows that (4.3) can be written as the ratio of the standard normal random variable $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j)$ over the square root of $\hat{\sigma}^2/\sigma^2$. These random variables can be shown to be independent, and $(n - k - 1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}$. The result then follows from the definition of a t random variable (see Section B.5).

Theorem 4.2 is important in that it allows us to test hypotheses involving the β_j. In most applications, our primary interest lies in testing the **null hypothesis**

$$H_0: \beta_j = 0, \tag{4.4}$$

where j corresponds to any of the k independent variables. It is important to understand what (4.4) means and to be able to describe this hypothesis in simple language for a particular application. Since β_j measures the partial effect of x_j on (the expected value of) y, after controlling for all other independent variables, (4.4) means that, once $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ have been accounted for, x_j has *no effect* on the expected value of y. We cannot state the null hypothesis as "x_j does have a partial effect on y" because this is true for any value of β_j other than zero. Classical testing is suited for testing *simple hypotheses* like (4.4).

As an example, consider the wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

The null hypothesis H_0: $\beta_2 = 0$ means that, once education and tenure have been accounted for, the number of years in the work force (*exper*) has no effect on hourly wage. This is an economically interesting hypothesis. If it is true, it implies that a person's work history prior to the current employment does not affect wage. If $\beta_2 > 0$, then prior work experience contributes to productivity, and hence to wage.

You probably remember from your statistics course the rudiments of hypothesis testing for the mean from a normal population. (This is reviewed in Appendix C.) The mechanics of testing (4.4) in the multiple regression context are very similar. The hard part is obtaining the coefficient estimates, the standard errors, and the critical values, but most of this work is done automatically by econometrics software. Our job is to learn how regression output can be used to test hypotheses of interest.

The statistic we use to test (4.4) (against any alternative) is called "the" **t statistic** or "the" **t ratio** of $\hat{\beta}_j$ and is defined as

$$t_{\hat{\beta}_j} \equiv \hat{\beta}_j/\text{se}(\hat{\beta}_j). \quad \textbf{(4.5)}$$

We have put "the" in quotation marks because, as we will see shortly, a more general form of the t statistic is needed for testing other hypotheses about β_j. For now, it is important to know that (4.5) is suitable only for testing (4.4). When it causes no confusion, we will sometimes write t in place of $t_{\hat{\beta}_j}$.

The t statistic for $\hat{\beta}_j$ is simple to compute given $\hat{\beta}_j$ and its standard error. In fact, most regression packages do the division for you and report the t statistic along with each coefficient and its standard error.

Before discussing how to use (4.5) formally to test H_0: $\beta_j = 0$, it is useful to see why $t_{\hat{\beta}_j}$ has features that make it reasonable as a test statistic to detect $\beta_j \neq 0$. First, since $\text{se}(\hat{\beta}_j)$ is always positive, $t_{\hat{\beta}_j}$ has the same sign as $\hat{\beta}_j$: if $\hat{\beta}_j$ is positive, then so is $t_{\hat{\beta}_j}$, and if $\hat{\beta}_j$ is negative, so is $t_{\hat{\beta}_j}$. Second, for a given value of $\text{se}(\hat{\beta}_j)$, a larger value of $\hat{\beta}_j$ leads to larger values of $t_{\hat{\beta}_j}$. If $\hat{\beta}_j$ becomes more negative, so does $t_{\hat{\beta}_j}$.

Since we are testing H_0: $\beta_j = 0$, it is only natural to look at our unbiased estimator of β_j, $\hat{\beta}_j$, for guidance. In any interesting application, the point estimate $\hat{\beta}_j$ will *never* exactly be zero, whether or not H_0 is true. The question is: How far is $\hat{\beta}_j$ from zero? A sample value of $\hat{\beta}_j$ very far from zero provides evidence against H_0: $\beta_j = 0$. However, we must recognize that there is a sampling error in our estimate $\hat{\beta}_j$, so the size of $\hat{\beta}_j$ must be weighed against its sampling error. Since the the standard error of $\hat{\beta}_j$ is an estimate of the standard deviation of $\hat{\beta}_j$, $t_{\hat{\beta}_j}$ measures how many estimated standard deviations $\hat{\beta}_j$ is away from zero. This is precisely what we do in testing whether the mean of a population is zero, using the standard t statistic from introductory statistics. Values of $t_{\hat{\beta}_j}$ sufficiently far from zero will result in a rejection of H_0. The precise rejection rule depends on the alternative hypothesis and the chosen significance level of the test.

Determining a rule for rejecting (4.4) at a given significance level—that is, the probability of rejecting H_0 when it is true—requires knowing the sampling distribution of $t_{\hat{\beta}_j}$ when H_0 is true. From Theorem 4.2, we know this to be t_{n-k-1}. This is the key theoretical result needed for testing (4.4).

Before proceeding, it is important to remember that we are testing hypotheses about the *population* parameters. We are *not* testing hypotheses about the estimates from a particular sample. Thus, it never makes sense to state a null hypothesis as "H_0: $\hat{\beta}_1 = 0$" or, even worse, as "H_0: .237 = 0" when the estimate of a parameter is .237 in the sample. We are testing whether the unknown population value, β_1, is zero.

Some treatments of regression analysis define the t statistic as the *absolute value* of (4.5), so that the t statistic is always positive. This practice has the drawback of making testing against one-sided alternatives clumsy. Throughout this text, the t statistic always has the same sign as the corresponding OLS coefficient estimate.

Testing Against One-Sided Alternatives

In order to determine a rule for rejecting H_0, we need to decide on the relevant **alternative hypothesis**. First consider a **one-sided alternative** of the form

$$H_1: \beta_j > 0. \quad \textbf{(4.6)}$$

This means that we do not care about alternatives to H_0 of the form H_1: $\beta_j < 0$; for some reason, perhaps on the basis of introspection or economic theory, we are ruling out population values of β_j less than zero. (Another way to think about this is that the null hypothesis is actually H_0: $\beta_j \leq 0$; in either case, the statistic $t_{\hat{\beta}_j}$ is used as the test statistic.)

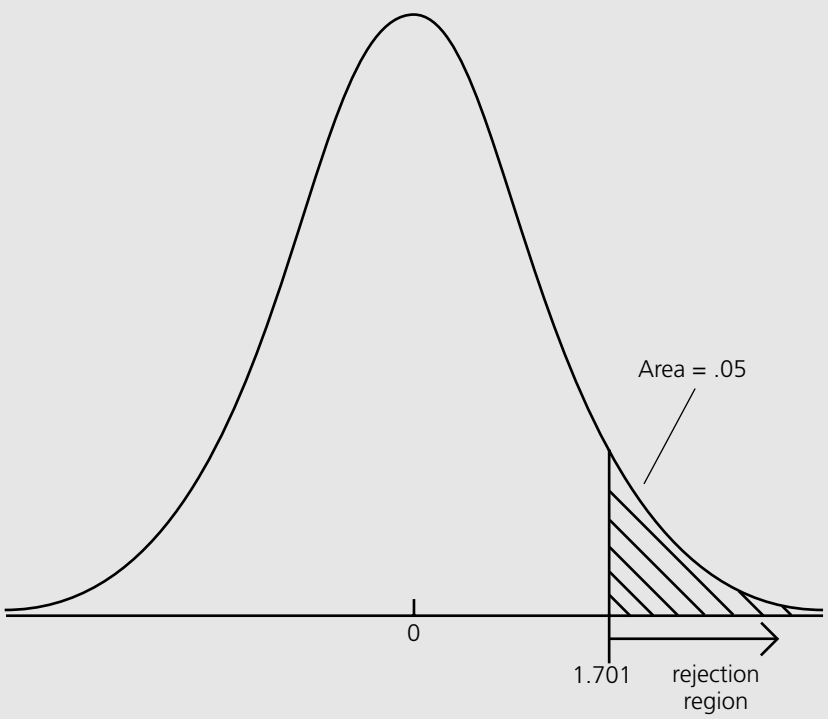
How should we choose a rejection rule? We must first decide on a **significance level** or the probability of rejecting H_0 when it is in fact true. For concreteness, suppose we have decided on a 5% significance level, as this is the most popular choice. Thus, we are willing to mistakenly reject H_0 when it is true 5% of the time. Now, while $t_{\hat{\beta}_j}$ has a t distribution under H_0—so that it has zero mean—under the alternative $\beta_j > 0$, the expected value of $t_{\hat{\beta}_j}$ is positive. Thus, we are looking for a "sufficiently large" positive value of $t_{\hat{\beta}_j}$ in order to reject H_0: $\beta_j = 0$ in favor of H_1: $\beta_j > 0$. Negative values of $t_{\hat{\beta}_j}$ provide no evidence in favor of H_1.

The definition of "sufficiently large," with a 5% significance level, is the 95th percentile in a t distribution with $n - k - 1$ degrees of freedom; denote this by c. In other words, the **rejection rule** is that H_0 is rejected in favor of H_1 at the 5% significance level if

$$t_{\hat{\beta}_j} > c. \tag{4.7}$$

Figure 4.2

5% rejection rule for the alternative H_1: $\beta_j > 0$ with 28 *df*.



By our choice of the **critical value** c, rejection of H_0 will occur for 5% of all random samples when H_0 is true.

The rejection rule in (4.7) is an example of a **one-tailed test**. In order to obtain c, we only need the significance level and the degrees of freedom. For example, for a 5% level test and with $n - k - 1 = 28$ degrees of freedom, the critical value is $c = 1.701$. If $t_{\hat{\beta}_j} < 1.701$, then we fail to reject H_0 in favor of (4.6) at the 5% level. Note that a negative value for $t_{\hat{\beta}_j}$, no matter how large in absolute value, leads to a failure in rejecting H_0 in favor of (4.6). (See Figure 4.2.)

The same procedure can be used with other significance levels. For a 10% level test and if $df = 21$, the critical value is $c = 1.323$. For a 1% significance level and if $df = 21$, $c = 2.518$. All of these critical values are obtained directly from Table G.2. You should note a pattern in the critical values: as the significance level falls, the critical value increases, so that we require a larger and larger value of $t_{\hat{\beta}_j}$ in order to reject H_0. Thus, if H_0 is rejected at, say, the 5% level, then it is automatically rejected at the 10% level as well. It makes no sense to reject the null hypothesis at, say, the 5% level and then to redo the test to determine the outcome at the 10% level.

As the degrees of freedom in the t distribution get large, the t distribution approaches the standard normal distribution. For example, when $n - k - 1 = 120$, the 5% critical value for the one-sided alternative (4.7) is 1.658, compared with the standard normal value of 1.645. These are close enough for practical purposes; for degrees of freedom greater than 120, one can use the standard normal critical values.

EXAMPLE 4.1
(Hourly Wage Equation)

Using the data in WAGE1.RAW gives the estimated equation

$$\widehat{\log(wage)} = \underset{(.104)}{.284} + \underset{(.007)}{.092}\, educ + \underset{(.0017)}{.0041}\, exper + \underset{(.003)}{.022}\, tenure$$
$$n = 526,\ R^2 = .316,$$

where standard errors appear in parentheses below the estimated coefficients. We will follow this convention throughout the text. This equation can be used to test whether the return to *exper*, controlling for *educ* and *tenure*, is zero in the population, against the alternative that it is positive. Write this as H_0: $\beta_{exper} = 0$ versus H_1: $\beta_{exper} > 0$. (In applications, indexing a parameter by its associated variable name is a nice way to label parameters, since the numerical indices that we use in the general model are arbitrary and can cause confusion.) Remember that β_{exper} denotes the unknown population parameter. It is nonsense to write "H_0: .0041 = 0" or "H_0: $\hat{\beta}_{exper} = 0$."

Since we have 522 degrees of freedom, we can use the standard normal critical values. The 5% critical value is 1.645, and the 1% critical value is 2.326. The t statistic for $\hat{\beta}_{exper}$ is

$$t_{\hat{\beta}_{exper}} = .0041/.0017 \approx 2.41,$$

and so $\hat{\beta}_{exper}$, or *exper*, is statistically significant even at the 1% level. We also say that "$\hat{\beta}_{exper}$ is statistically greater than zero at the 1% significance level."

The estimated return for another year of experience, holding tenure and education fixed, is not large. For example, adding three more years increases log(*wage*) by 3(.0041) =

.0123, so wage is only about 1.2% higher. Nevertheless, we have persuasively shown that the partial effect of experience *is* positive in the population.

The one-sided alternative that the parameter is less than zero,

$$H_1: \beta_j < 0, \tag{4.8}$$

also arises in applications.

The rejection rule for alternative (4.8) is just the mirror image of the previous case. Now, the critical value comes from the left tail of the t distribution. In practice, it is easiest to think of the rejection rule as

$$t_{\hat{\beta}_j} < -c, \tag{4.9}$$

where c is the critical value for the alternative $H_1: \beta_j > 0$. For simplicity, we always assume c is positive, since this is how critical values are reported in t tables, and so the critical value $-c$ is a negative number.

For example, if the significance level is 5% and the degrees of freedom is 18, then $c = 1.734$, and so $H_0: \beta_j = 0$ is rejected in favor of $H_1: \beta_j < 0$ at the 5% level if $t_{\hat{\beta}_j} < -1.734$. It is important to remember that, to reject H_0 against the negative alternative (4.8), we must get a negative t statistic. A positive t ratio, no matter how large, provides no evidence in favor of (4.8). The rejection rule is illustrated in Figure 4.3.

QUESTION 4.2

Let community loan approval rates be determined by

$$apprate = \beta_0 + \beta_1 percmin + \beta_2 avginc + \beta_3 avgwlth + \beta_4 avgdebt + u,$$

where *percmin* is the percent minority in the community, *avginc* is average income, *avgwlth* is average wealth, and *avgdebt* is some measure of average debt obligations. How do you state the null hypothesis that there is *no* difference in loan rates across neighborhoods due to racial and ethnic composition, when average income, average wealth, and average debt have been controlled for? How do you state the alternative that there is discrimination against minorities in loan approval rates?

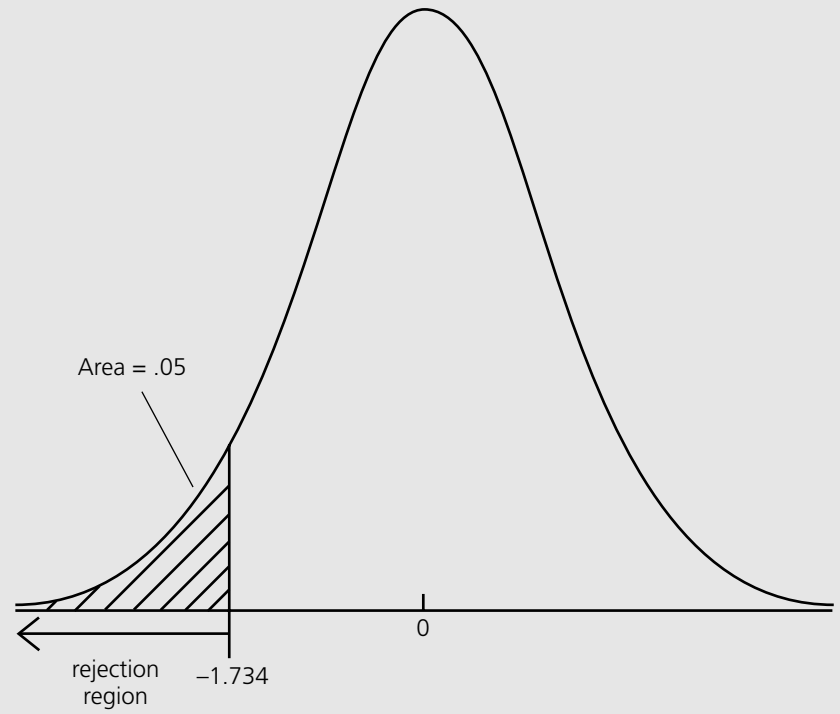
EXAMPLE 4.2
(Student Performance and School Size)

There is much interest in the effect of school size on student performance. (See, for example, *The New York Times Magazine*, 5/28/95.) One claim is that, everything else being equal, students at smaller schools fare better than those at larger schools. This hypothesis is assumed to be true even after accounting for differences in class sizes across schools.

The file MEAP93.RAW contains data on 408 high schools in Michigan for the year 1993. We can use these data to test the null hypothesis that school size has no effect on standardized test scores, against the alternative that size has a negative effect. Performance is measured by the percentage of students receiving a passing score on the Michigan Educational Assessment Program (MEAP) standardized tenth grade math test (*math10*). School size is measured by student enrollment (*enroll*). The null hypothesis is $H_0: \beta_{enroll} = 0$, and the alternative is $H_1: \beta_{enroll} < 0$. For now, we will control for two other factors, average annual teacher compensation (*totcomp*) and the number of staff per one thousand students (*staff*). Teacher compensation is a measure of teacher quality, and staff size is a rough measure of how much attention students receive.

Figure 4.3

5% rejection rule for the alternative H_1: $\beta_j < 0$ with 18 *df*.



The estimated equation, with standard errors in parentheses, is

$$\widehat{math10} = \underset{(6.113)}{2.274} + \underset{(.00010)}{.00046}\ totcomp + \underset{(.040)}{.048}\ staff - \underset{(.00022)}{.00020}\ enroll$$
$$n = 408, R^2 = .0541.$$

The coefficient on *enroll*, $-.0002$, is in accordance with the conjecture that larger schools hamper performance: higher enrollment leads to a lower percentage of students with a passing tenth grade math score. (The coefficients on *totcomp* and *staff* also have the signs we expect.) The fact that *enroll* has an estimated coefficient different from zero could just be due to sampling error; to be convinced of an effect, we need to conduct a *t* test.

Since $n - k - 1 = 408 - 4 = 404$, we use the standard normal critical value. At the 5% level, the critical value is -1.65; the *t* statistic on *enroll* must be *less* than -1.65 to reject H_0 at the 5% level.

The *t* statistic on *enroll* is $-.0002/.00022 \approx -.91$, which is larger than -1.65: we *fail* to reject H_0 in favor of H_1 at the 5% level. In fact, the 15% critical value is -1.04, and since $-.91 > -1.04$, we fail to reject H_0 even at the 15% level. We conclude that *enroll* is not statistically significant at the 15% level.

The variable *totcomp* is statistically significant even at the 1% significance level because its *t* statistic is 4.6. On the other hand, the *t* statistic for *staff* is 1.2, and so we cannot reject H_0: $\beta_{staff} = 0$ against H_1: $\beta_{staff} > 0$ even at the 10% significance level. (The critical value is $c = 1.28$ from the standard normal distribution.)

To illustrate how changing functional form can affect our conclusions, we also estimate the model with all independent variables in logarithmic form. This allows, for example, the school size effect to diminish as school size increases. The estimated equation is

$$\widehat{math10} = \underset{(48.70)}{-207.66} + \underset{(4.06)}{21.16}\ \log(totcomp) + \underset{(4.19)}{3.98}\ \log(staff) - \underset{(0.69)}{1.29}\ \log(enroll)$$
$$n = 408,\ R^2 = .0654.$$

The *t* statistic on log(*enroll*) is about -1.87; since this is below the 5% critical value -1.65, we reject H_0: $\beta_{\log(enroll)} = 0$ in favor of H_1: $\beta_{\log(enroll)} < 0$ at the 5% level.

In Chapter 2, we encountered a model where the dependent variable appeared in its original form (called *level* form), while the independent variable appeared in log form (called *level-log* model). The interpretation of the parameters is the same in the multiple regression context, except, of course, that we can give the parameters a ceteris paribus interpretation. Holding *totcomp* and *staff* fixed, we have $\Delta\widehat{math10} = -1.29[\Delta\log(enroll)]$, so that

$$\Delta\widehat{math10} \approx -(1.29/100)(\%\Delta enroll) \approx -.013(\%\Delta enroll).$$

Once again, we have used the fact that the change in log(*enroll*), when multiplied by 100, is approximately the percentage change in *enroll*. Thus, if enrollment is 10% higher at a school, $\widehat{math10}$ is predicted to be 1.3 percentage points lower (*math10* is measured as a percent).

Which model do we prefer: the one using the level of *enroll* or the one using log(*enroll*)? In the level-level model, enrollment does not have a statistically significant effect, but in the level-log model it does. This translates into a higher *R*-squared for the level-log model, which means we explain more of the variation in *math10* by using *enroll* in logarithmic form (6.5% to 5.4%). The level-log model is preferred, as it more closely captures the relationship between *math10* and *enroll*. We will say more about using *R*-squared to choose functional form in Chapter 6.

Two-Sided Alternatives

In applications, it is common to test the null hypothesis H_0: $\beta_j = 0$ against a **two-sided alternative**, that is,

$$H_1: \beta_j \neq 0. \qquad \textbf{(4.10)}$$

Under this alternative, x_j has a ceteris paribus effect on y without specifying whether the effect is positive or negative. This is the relevant alternative when the sign of β_j is not well-determined by theory (or common sense). Even when we know whether β_j is positive or negative under the alternative, a two-sided test is often prudent. At a minimum,

using a two-sided alternative prevents us from looking at the estimated equation and then basing the alternative on whether $\hat{\beta}_j$ is positive or negative. Using the regression estimates to help us formulate the null or alternative hypotheses is not allowed because classical statistical inference presumes that we state the null and alternative about the population before looking at the data. For example, we should not first estimate the equation relating math performance to enrollment, note that the estimated effect is negative, and then decide the relevant alternative is H_1: $\beta_{enroll} < 0$.

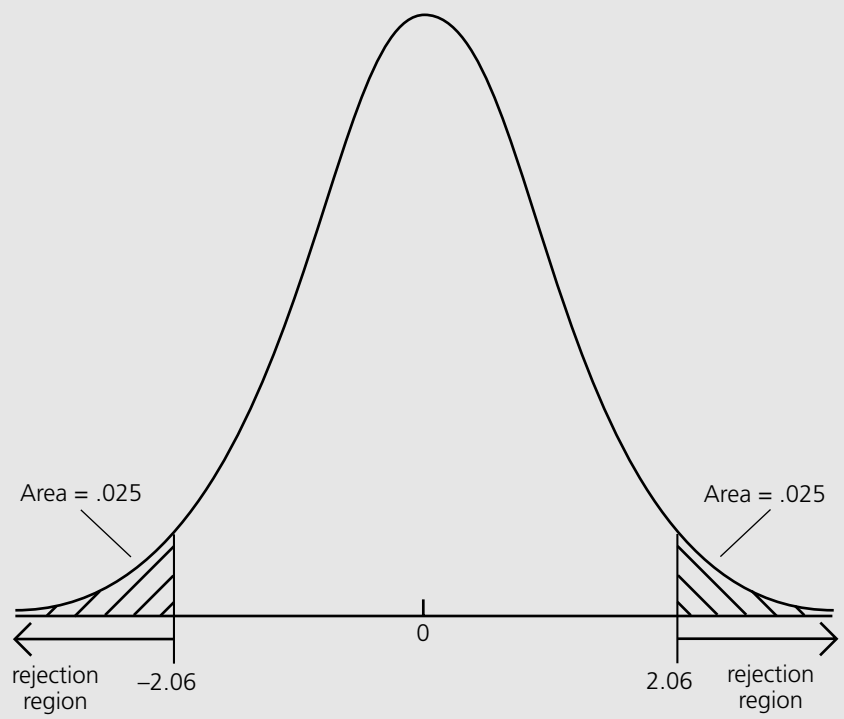
When the alternative is two-sided, we are interested in the *absolute value* of the t statistic. The rejection rule for H_0: $\beta_j = 0$ against (4.10) is

$$|t_{\hat{\beta}_j}| > c, \tag{4.11}$$

where $|\cdot|$ denotes absolute value and c is an appropriately chosen critical value. To find c, we again specify a significance level, say 5%. For a **two-tailed test**, c is chosen to make the area in each tail of the t distribution equal 2.5%. In other words, c is the 97.5th percentile in the t distribution with $n - k - 1$ degrees of freedom. When $n - k - 1 = 25$, the 5% critical value for a two-sided test is $c = 2.060$. Figure 4.4 provides an illustration of this distribution.

Figure 4.4

5% rejection rule for the alternative H_1: $\beta_j \neq 0$ with 25 *df*.



When a specific alternative is not stated, it is usually considered to be two-sided. In the remainder of this text, the default will be a two-sided alternative, and 5% will be the default significance level. When carrying out empirical econometric analysis, it is always a good idea to be explicit about the alternative and the significance level. If H_0 is rejected in favor of (4.10) at the 5% level, we usually say that "x_j is **statistically significant**, or statistically different from zero, at the 5% level." If H_0 is not rejected, we say that "x_j is **statistically insignificant** at the 5% level."

EXAMPLE 4.3
(Determinants of College GPA)

We use GPA1.RAW to estimate a model explaining college GPA (*colGPA*), with the average number of lectures missed per week (*skipped*) as an additional explanatory variable. The estimated model is

$$\widehat{colGPA} = \underset{(0.33)}{1.39} + \underset{(.094)}{.412}\, hsGPA + \underset{(.011)}{.015}\, ACT - \underset{(.026)}{.083}\, skipped$$
$$n = 141,\ R^2 = .234.$$

We can easily compute *t* statistics to see which variables are statistically significant, using a two-sided alternative in each case. The 5% critical value is about 1.96, since the degrees of freedom (141 − 4 = 137) is large enough to use the standard normal approximation. The 1% critical value is about 2.58.

The *t* statistic on *hsGPA* is 4.38, which is significant at very small significance levels. Thus, we say that "*hsGPA* is statistically significant at any *conventional* significance level." The *t* statistic on *ACT* is 1.36, which is not statistically significant at the 10% level against a two-sided alternative. The coefficient on *ACT* is also practically small: a 10-point increase in *ACT*, which is large, is predicted to increase *colGPA* by only .15 point. Thus, the variable *ACT* is practically, as well as statistically, insignificant.

The coefficient on *skipped* has a *t* statistic of $-.083/.026 = -3.19$, so *skipped* is statistically significant at the 1% significance level ($3.19 > 2.58$). This coefficient means that another lecture missed per week lowers predicted *colGPA* by about .083. Thus, holding *hsGPA* and *ACT* fixed, the predicted difference in *colGPA* between a student who misses no lectures per week and a student who misses five lectures per week is about .42. Remember that this says nothing about specific students, but pertains to average students across the population.

In this example, for each variable in the model, we could argue that a one-sided alternative is appropriate. The variables *hsGPA* and *skipped* are very significant using a two-tailed test and have the signs that we expect, so there is no reason to do a one-tailed test. On the other hand, against a one-sided alternative ($\beta_3 > 0$), *ACT* is significant at the 10% level but not at the 5% level. This does not change the fact that the coefficient on *ACT* is pretty small.

Testing Other Hypotheses About β_j

Although H_0: $\beta_j = 0$ is the most common hypothesis, we sometimes want to test whether β_j is equal to some other given constant. Two common examples are $\beta_j = 1$ and $\beta_j = -1$. Generally, if the null is stated as

$$H_0: \beta_j = a_j, \tag{4.12}$$

where a_j is our hypothesized value of β_j, then the appropriate t statistic is

$$t = (\hat{\beta}_j - a_j)/\text{se}(\hat{\beta}_j).$$

As before, t measures how many estimated standard deviations $\hat{\beta}_j$ is from the hypothesized value of β_j. The general t statistic is usefully written as

$$t = \frac{(\textit{estimate} - \textit{hypothesized value})}{\textit{standard error}}. \tag{4.13}$$

Under (4.12), this t statistic is distributed as t_{n-k-1} from Theorem 4.2. The usual t statistic is obtained when $a_j = 0$.

We can use the general t statistic to test against one-sided or two-sided alternatives. For example, if the null and alternative hypotheses are H_0: $\beta_j = 1$ and H_1: $\beta_j > 1$, then we find the critical value for a one-sided alternative *exactly* as before: the difference is in how we compute the t statistic, not in how we obtain the appropriate c. We reject H_0 in favor of H_1 if $t > c$. In this case, we would say that "$\hat{\beta}_j$ is statistically greater than one" at the appropriate significance level.

EXAMPLE 4.4
(Campus Crime and Enrollment)

Consider a simple model relating the annual number of crimes on college campuses (*crime*) to student enrollment (*enroll*):

$$\log(\textit{crime}) = \beta_0 + \beta_1 \log(\textit{enroll}) + u.$$

This is a constant elasticity model, where β_1 is the elasticity of crime with respect to enrollment. It is not much use to test H_0: $\beta_1 = 0$, as we expect the total number of crimes to increase as the size of the campus increases. A more interesting hypothesis to test would be that the elasticity of crime with respect to enrollment is one: H_0: $\beta_1 = 1$. This means that a 1% increase in enrollment leads to, on average, a 1% increase in crime. A noteworthy alternative is H_1: $\beta_1 > 1$, which implies that a 1% increase in enrollment increases campus crime by *more* than 1%. If $\beta_1 > 1$, then, in a relative sense—not just an absolute sense—crime is more of a problem on larger campuses. One way to see this is to take the exponential of the equation:

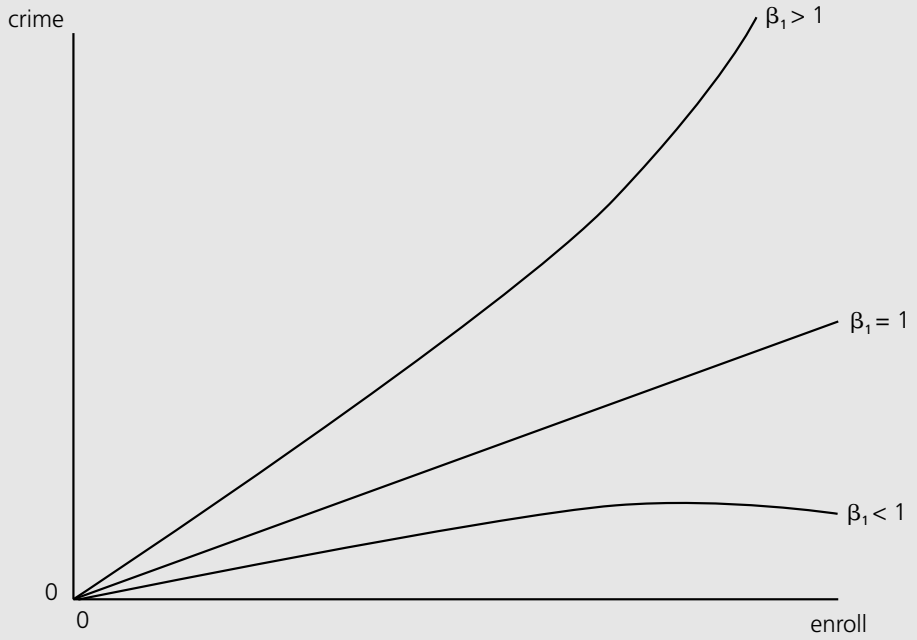
$$\textit{crime} = \exp(\beta_0)\textit{enroll}^{\beta_1}\exp(u).$$

(See Appendix A for properties of the natural logarithm and exponential functions.) For $\beta_0 = 0$ and $u = 0$, this equation is graphed in Figure 4.5 for $\beta_1 < 1$, $\beta_1 = 1$, and $\beta_1 > 1$.

We test $\beta_1 = 1$ against $\beta_1 > 1$ using data on 97 colleges and universities in the United States for the year 1992. The data come from the FBI's *Uniform Crime Reports*, and the average number of campus crimes in the sample is about 394, while the average enrollment is about 16,076. The estimated equation (with estimates and standard errors rounded to two decimal places) is

Figure 4.5

Graph of *crime* = *enroll*$^{\beta_1}$ for $\beta_1 < 1$, $\beta_1 = 1$, and $\beta_1 > 1$.



$$\widehat{\log(crime)} = -6.63 + 1.27 \log(enroll) \quad \textbf{(4.14)}$$
$$(1.03) \quad (0.11)$$
$$n = 97, R^2 = .585.$$

The estimated elasticity of *crime* with respect to *enroll*, 1.27, is in the direction of the alternative $\beta_1 > 1$. But is there enough evidence to conclude that $\beta_1 > 1$? We need to be careful in testing this hypothesis, especially because the statistical output of standard regression packages is much more complex than the simplified output reported in equation (4.14). Our first instinct might be to construct "the" *t* statistic by taking the coefficient on log(*enroll*) and dividing it by its standard error, which is the *t* statistic reported by a regression package. But this is the *wrong* statistic for testing H_0: $\beta_1 = 1$. The correct *t* statistic is obtained from (4.13): we subtract the hypothesized value, unity, from the estimate and divide the result by the standard error of $\hat{\beta}_1$: $t = (1.27 - 1)/.11 = .27/.11 \approx 2.45$. The one-sided 5% critical value for a *t* distribution with $97 - 2 = 95$ *df* is about 1.66 (using $df = 120$), so we clearly reject $\beta_1 = 1$ in favor of $\beta_1 > 1$ at the 5% level. In fact, the 1% critical value is about 2.37, and so we reject the null in favor of the alternative at even the 1% level.

We should keep in mind that this analysis holds no other factors constant, so the elasticity of 1.27 is not necessarily a good estimate of ceteris paribus effect. It could be that

larger enrollments are correlated with other factors that cause higher crime: larger schools might be located in higher crime areas. We could control for this by collecting data on crime rates in the local city.

For a two-sided alternative, for example H_0: $\beta_j = -1$, H_1: $\beta_j \neq -1$, we still compute the t statistic as in (4.13): $t = (\hat{\beta}_j + 1)/\text{se}(\hat{\beta}_j)$ (notice how subtracting -1 means adding 1). The rejection rule is the usual one for a two-sided test: reject H_0 if $|t| > c$, where c is a two-tailed critical value. If H_0 is rejected, we say that "$\hat{\beta}_j$ is statistically different from negative one" at the appropriate significance level.

EXAMPLE 4.5
(Housing Prices and Air Pollution)

For a sample of 506 communities in the Boston area, we estimate a model relating median housing price (*price*) in the community to various community characteristics: *nox* is the amount of nitrous oxide in the air, in parts per million; *dist* is a weighted distance of the community from five employment centers, in miles; *rooms* is the average number of rooms in houses in the community; and *stratio* is the average student-teacher ratio of schools in the community. The population model is

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 stratio + u.$$

Thus, β_1 is the elasticity of *price* with respect to *nox*. We wish to test H_0: $\beta_1 = -1$ against the alternative H_1: $\beta_1 \neq -1$. The t statistic for doing this test is $t = (\hat{\beta}_1 + 1)/\text{se}(\hat{\beta}_1)$.

Using the data in HPRICE2.RAW, the estimated model is

$$\widehat{\log(price)} = \underset{(0.32)}{11.08} - \underset{(.117)}{.954}\log(nox) - \underset{(.043)}{.134}\log(dist) + \underset{(.019)}{.255}\,rooms - \underset{(.006)}{.052}\,stratio$$

$$n = 506,\ R^2 = .581.$$

The slope estimates all have the anticipated signs. Each coefficient is statistically different from zero at very small significance levels, including the coefficient on log(*nox*). But we do not want to test that $\beta_1 = 0$. The null hypothesis of interest is H_0: $\beta_1 = -1$, with corresponding t statistic $(-.954 + 1)/.117 = .393$. There is little need to look in the t table for a critical value when the t statistic is this small: the estimated elasticity is not statistically different from -1 even at very large significance levels. Controlling for the factors we have included, there is little evidence that the elasticity is different from -1.

Computing *p*-values for *t* tests

So far, we have talked about how to test hypotheses using a classical approach: after stating the alternative hypothesis, we choose a significance level, which then determines a critical value. Once the critical value has been identified, the value of the t statistic is compared with the critical value, and the null is either rejected or not rejected at the given significance level.

Even after deciding on the appropriate alternative, there is a component of arbitrariness to the classical approach, which results from having to choose a significance level ahead of time. Different researchers prefer different significance levels, depending on the particular application. There is no "correct" significance level.

Committing to a significance level ahead of time can hide useful information about the outcome of a hypothesis test. For example, suppose that we wish to test the null hypothesis that a parameter is zero against a two-sided alternative, and with 40 degrees of freedom we obtain a t statistic equal to 1.85. The null hypothesis is not rejected at the 5% level, since the t statistic is less than the two-tailed critical value of $c = 2.021$. A researcher whose agenda is not to reject the null could simply report this outcome along with the estimate: the null hypothesis is not rejected at the 5% level. Of course, if the t statistic, or the coefficient and its standard error, are reported, then we can also determine that the null hypothesis would be rejected at the 10% level, since the 10% critical value is $c = 1.684$.

Rather than testing at different significance levels, it is more informative to answer the following question: Given the observed value of the t statistic, what is the *smallest* significance level at which the null hypothesis would be rejected? This level is known as the ***p*-value** for the test (see Appendix C). In the previous example, we know the p-value is greater than .05, since the null is not rejected at the 5% level, and we know that the p-value is less than .10, since the null is rejected at the 10% level. We obtain the actual p-value by computing the probability that a t random variable, with 40 df, is larger than 1.85 in absolute value. That is, the p-value is the significance level of the test when we use the value of the test statistic, 1.85 in the above example, as the critical value for the test. This p-value is shown in Figure 4.6.

Since a p-value is a probability, its value is always between zero and one. In order to compute p-values, we either need extremely detailed printed tables of the t distribution—which is not very practical—or a computer program that computes areas under the probability density function of the t distribution. Most modern regression packages have this capability. Some packages compute p-values routinely with each OLS regression, but only for certain hypotheses. If a regression package reports a p-value along with the standard OLS output, it is almost certainly the p-value for testing the null hypothesis H_0: $\beta_j = 0$ against the two-sided alternative. The p-value in this case is

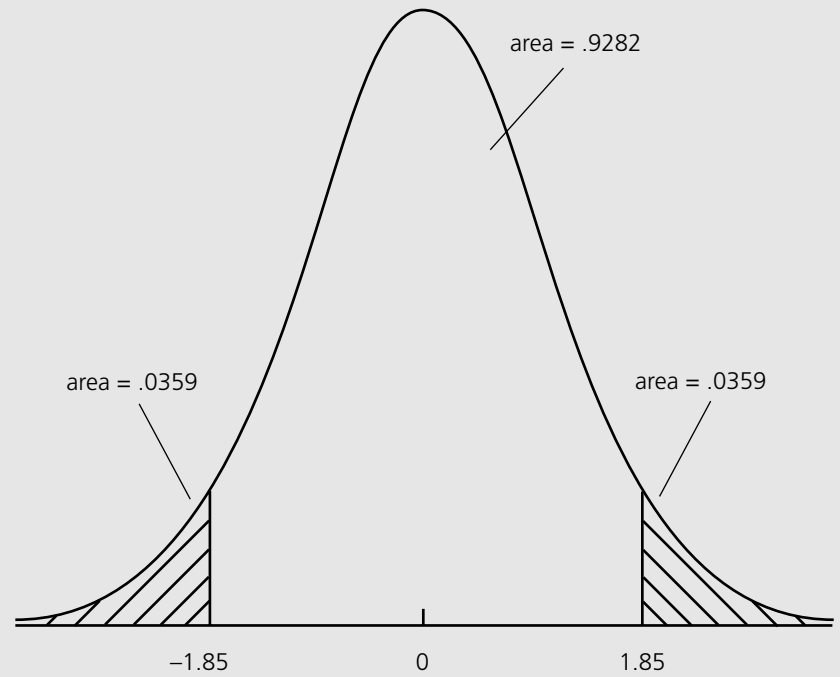
$$P(|T| > |t|), \tag{4.15}$$

where, for clarity, we let T denote a t distributed random variable with $n - k - 1$ degrees of freedom and let t denote the numerical value of the test statistic.

The p-value nicely summarizes the strength or weakness of the empirical evidence against the null hypothesis. Perhaps its most useful interpretation is the following: the p-value is the probability of observing a t statistic as extreme as we did *if the null hypothesis is true*. This means that *small* p-values are evidence *against* the null; large p-values provide little evidence against H_0. For example, if the p-value = .50 (reported always as a decimal, not a percent), then we would observe a value of the t statistic as extreme as we did in 50% of all random samples when the null hypothesis is true; this is pretty weak evidence against H_0.

Figure 4.6

Obtaining the *p*-value against a two-sided alternative, when $t = 1.85$ and $df = 40$.



In the example with $df = 40$ and $t = 1.85$, the *p*-value is computed as

$$p\text{-value} = P(|T| > 1.85) = 2P(T > 1.85) = 2(.0359) = .0718,$$

where $P(T > 1.85)$ is the area to the right of 1.85 in a *t* distribution with 40 *df*. (This value was computed using the econometrics package Stata; it is not available in Table G.2.) This means that, if the null hypothesis is true, we would observe an absolute value of the *t* statistic as large as 1.85 about 7.2% of the time. This provides some evidence against the null hypothesis, but we would not reject the null at the 5% significance level.

The previous example illustrates that once the *p*-value has been computed, a classical test can be carried out at any desired level. If α denotes the significance level of the test (in decimal form), then H_0 is rejected if $p\text{-value} < \alpha$; otherwise H_0 is not rejected at the $100 \cdot \alpha\%$ level.

Computing *p*-values for one-sided alternatives is also quite simple. Suppose, for example, that we test $H_0: \beta_j = 0$ against $H_1: \beta_j > 0$. If $\hat{\beta}_j < 0$, then computing a *p*-value is not important: we know that the *p*-value is greater than .50, which will never cause us to reject H_0 in favor of H_1. If $\hat{\beta}_j > 0$, then $t > 0$ and the *p*-value is just the probability that a *t* random variable with the appropriate *df* exceeds the value *t*. Some regression packages only compute *p*-values for two-sided alternatives. But it is simple to obtain the one-sided *p*-value: just divide the two-sided *p*-value by 2.

If the alternative is H_1: $\beta_j < 0$, it makes sense to compute a *p*-value if $\hat{\beta}_j < 0$ (and hence $t < 0$): *p*-value $= P(T < t) = P(T > |t|)$ because the *t* distribution is symmetric about zero. Again, this can be obtained as one-half of the *p*-value for the two-tailed test.

QUESTION 4.3

Suppose you estimate a regression model and obtain $\hat{\beta}_1 = .56$ and *p*-value = .086 for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. What is the *p*-value for testing H_0: $\beta_1 = 0$ against H_1: $\beta_1 > 0$?

Because you will quickly become familiar with the magnitudes of *t* statistics that lead to statistical significance, especially for large sample sizes, it is not always crucial to report *p*-values for *t* statistics. But it does not hurt to report them. Further, when we discuss *F* testing in Section 4.5, we will see that it is important to compute *p*-values, because critical values for *F* tests are not so easily memorized.

A Reminder on the Language of Classical Hypothesis Testing

When H_0 is not rejected, we prefer to use the language "we fail to reject H_0 at the x% level," rather than "H_0 is accepted at the x% level." We can use Example 4.5 to illustrate why the former statement is preferred. In this example, the estimated elasticity of *price* with respect to *nox* is −.954, and the *t* statistic for testing H_0: $\beta_{nox} = -1$ is $t = .393$; therefore, we cannot reject H_0. But there are many other values for β_{nox} (more than we can count) that cannot be rejected. For example, the *t* statistic for H_0: $\beta_{nox} = -.9$ is $(-.954 + .9)/.117 = -.462$, and so this null is not rejected either. Clearly $\beta_{nox} = -1$ and $\beta_{nox} = -.9$ cannot both be true, so it makes no sense to say that we "accept" either of these hypotheses. All we can say is that the data do not allow us to reject either of these hypotheses at the 5% significance level.

Economic, or Practical, versus Statistical Significance

Since we have emphasized *statistical significance* throughout this section, now is a good time to remember that we should pay attention to the magnitude of the *coefficient* estimates in addition to the size of the *t* statistics. The statistical significance of a variable x_j is determined entirely by the size of $t_{\hat{\beta}_j}$, whereas the **economic significance** or **practical significance** of a variable is related to the size (and sign) of $\hat{\beta}_j$.

Recall that the *t* statistic for testing H_0: $\beta_j = 0$ is defined by dividing the estimate by its standard error: $t_{\hat{\beta}_j} = \hat{\beta}_j/\text{se}(\hat{\beta}_j)$. Thus, $t_{\hat{\beta}_j}$ can indicate statistical significance either because $\hat{\beta}_j$ is "large" or because $\text{se}(\hat{\beta}_j)$ is "small." It is important in practice to distinguish between these reasons for statistically significant *t* statistics. Too much focus on statistical significance can lead to the false conclusion that a variable is "important" for explaining *y* even though its estimated effect is modest.

EXAMPLE 4.6
[Participation Rates in 401(k) Plans]

In Example 3.3, we used the data on 401(k) plans to estimate a model describing participation rates in terms of the firm's match rate and the age of the plan. We now include a measure of firm size, the total number of firm employees (*totemp*). The estimated equation is

$$\widehat{prate} = 80.29 + 5.44\ mrate + .269\ age - .00013\ totemp$$
$$(0.78)\quad (0.52)\quad (.045)\quad (.00004)$$
$$n = 1{,}534,\ R^2 = .100.$$

The smallest *t* statistic in absolute value is that on the variable *totemp*: $t = -.00013/.00004 = -3.25$, and this is statistically significant at very small significance levels. (The two-tailed *p*-value for this *t* statistic is about .001.) Thus, all of the variables are statistically significant at rather small significance levels.

How big, in a practical sense, is the coefficient on *totemp*? Holding *mrate* and *age* fixed, if a firm grows by 10,000 employees, the participation rate falls by 10,000(.00013) = 1.3 percentage points. This is a huge increase in number of employees with only a modest effect on the participation rate. Thus, while firm size does affect the participation rate, the effect is not practically very large.

The previous example shows that it is especially important to interpret the magnitude of the coefficient, in addition to looking at *t* statistics, when working with large samples. With large sample sizes, parameters can be estimated very precisely: standard errors are often quite small relative to the coefficient estimates, which usually results in statistical significance.

Some researchers insist on using smaller significance levels as the sample size increases, partly as a way to offset the fact that standard errors are getting smaller. For example, if we feel comfortable with a 5% level when *n* is a few hundred, we might use the 1% level when *n* is a few thousand. Using a smaller significance level means that economic and statistical significance are more likely to coincide, but there are no guarantees: in the the previous example, even if we use a significance level as small as .1% (one-tenth of one percent), we would still conclude that *totemp* is statistically significant.

Most researchers are also willing to entertain larger significance levels in applications with small sample sizes, reflecting the fact that it is harder to find significance with smaller sample sizes (the critical values are larger in magnitude and the estimators are less precise). Unfortunately, whether or not this is the case can depend on the researcher's underlying agenda.

EXAMPLE 4.7
(Effect of Job Training Grants on Firm Scrap Rates)

The scrap rate for a manufacturing firm is the number of defective items out of every 100 items produced that must be discarded. Thus, a decrease in the scrap rate reflects higher productivity.

We can use the scrap rate to measure the effect of worker training on productivity. For a sample of Michigan manufacturing firms in 1987, the following equation is estimated:

$$\widehat{\log(scrap)} = 13.72 - .028\ hrsemp - 1.21\ \log(sales) + 1.48\ \log(employ)$$
$$(4.91)\quad (.019)\quad (0.41)\quad (0.43)$$
$$n = 30,\ R^2 = .431.$$

(This regression uses a subset of the data in JTRAIN.RAW.) The variable *hrsemp* is annual hours of training per employee, *sales* is annual firm sales (in dollars), and *employ* is number of firm employees. The average scrap rate in the sample is about 3.5, and the average *hrsemp* is about 7.3.

The main variable of interest is *hrsemp*. One more hour of training per employee lowers log(*scrap*) by .028, which means the scrap rate is about 2.8% lower. Thus, if *hrsemp* increases by 5—each employee is trained 5 more hours per year—the scrap rate is estimated to fall by 5(2.8) = 14%. This seems like a reasonably large effect, but whether the additional training is worthwhile to the firm depends on the cost of training and the benefits from a lower scrap rate. We do not have the numbers needed to do a cost benefit analysis, but the estimated effect seems nontrivial.

What about the *statistical significance* of the training variable? The *t* statistic on *hrsemp* is $-.028/.019 = -1.47$, and now you probably recognize this as not being large enough in magnitude to conclude that *hrsemp* is statistically significant at the 5% level. In fact, with $30 - 4 = 26$ degrees of freedom for the one-sided alternative H_1: $\beta_{hrsemp} < 0$, the 5% critical value is about -1.71. Thus, using a strict 5% level test, we must conclude that *hrsemp* is not statistically significant, even using a one-sided alternative.

Because the sample size is pretty small, we might be more liberal with the significance level. The 10% critical value is -1.32, and so *hrsemp* is significant against the one-sided alternative at the 10% level. The *p*-value is easily computed as $P(T_{26} < -1.47) = .077$. This may be a low enough *p*-value to conclude that the estimated effect of training is not just due to sampling error, but some economists would have different opinions on this.

Remember that large standard errors can also be a result of multicollinearity (high correlation among some of the independent variables), even if the sample size seems fairly large. As we discussed in Section 3.4, there is not much we can do about this problem other than to collect more data or change the scope of the analysis by dropping certain independent variables from the model. As in the case of a small sample size, it can be hard to precisely estimate partial effects when some of the explanatory variables are highly correlated. (Section 4.5 contains an example.)

We end this section with some guidelines for discussing the economic and statistical significance of a variable in a multiple regression model:

1. Check for statistical significance. If the variable is statistically significant, discuss the magnitude of the coefficient to get an idea of its practical or economic importance. This latter step can require some care, depending on how the independent and dependent variables appear in the equation. (In particular, what are the units of measurement? Do the variables appear in logarithmic form?)
2. If a variable is not statistically significant at the usual levels (10%, 5% or 1%), you might still ask if the variable has the expected effect on *y* and whether that effect is practically large. If it is large, you should compute a *p*-value for the *t* statistic. For small sample sizes, you can sometimes make a case for *p*-values as large as .20 (but there are no hard rules). With large *p*-values, that is, small *t* statistics, we are treading on thin ice because the practically large estimates may be due to sampling error: a different random sample could result in a very different estimate.

3. It is common to find variables with small t statistics that have the "wrong" sign. For practical purposes, these can be ignored: we conclude that the variables are statistically insignificant. A significant variable that has the unexpected sign and a practically large effect is much more troubling and difficult to resolve. One must usually think more about the model and the nature of the data in order to solve such problems. Often a counterintuitive, significant estimate results from the omission of a key variable or from one of the important problems we will discuss in Chapters 9 and 15.

4.3 CONFIDENCE INTERVALS

Under the classical linear model assumptions, we can easily construct a **confidence interval (CI)** for the population parameter β_j. Confidence intervals are also called *interval estimates* because they provide a range of likely values for the population parameter, and not just a point estimate.

Using the fact that $(\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j)$ has a t distribution with $n - k - 1$ degrees of freedom [see (4.3)], simple manipulation leads to a CI for the unknown β_j. A *95% confidence interval*, given by

$$\hat{\beta}_j \pm c \cdot \text{se}(\hat{\beta}_j), \tag{4.16}$$

where the constant c is the 97.5th percentile in a t_{n-k-1} distribution. More precisely, the lower and upper bounds of the confidence interval are given by

$$\underline{\beta}_j \equiv \hat{\beta}_j - c \cdot \text{se}(\hat{\beta}_j)$$

and

$$\bar{\beta}_j \equiv \hat{\beta}_j + c \cdot \text{se}(\hat{\beta}_j),$$

respectively.

At this point, it is useful to review the meaning of a confidence interval. If random samples were obtained over and over again, with $\underline{\beta}_j$, and $\bar{\beta}_j$ computed each time, then the (unknown) population value β_j would lie in the interval $(\underline{\beta}_j, \bar{\beta}_j)$ for 95% of the samples. Unfortunately, for the single sample that we use to contruct the CI, we do not know whether β_j is actually contained in the interval. We hope we have obtained a sample that is one of the 95% of all samples where the interval estimate contains β_j, but we have no guarantee.

Constructing a confidence interval is very simple when using current computing technology. Three quantities are needed: $\hat{\beta}_j$, $\text{se}(\hat{\beta}_j)$, and c. The coefficient estimate and its standard error are reported by any regression package. To obtain the value c, we must know the degrees of freedom, $n - k - 1$, and the level of confidence—95% in this case. Then, the value for c is obtained from the t_{n-k-1} distribution.

As an example, for $df = n - k - 1 = 25$, a 95% confidence interval for any β_j is given by $[\hat{\beta}_j - 2.06 \cdot \text{se}(\hat{\beta}_j), \hat{\beta}_j + 2.06 \cdot \text{se}(\hat{\beta}_j)]$.

When $n - k - 1 > 120$, the t_{n-k-1} distribution is close enough to normal to use the 97.5th percentile in a standard normal distribution for constructing a 95% CI: $\hat{\beta}_j \pm 1.96 \cdot \text{se}(\hat{\beta}_j)$. In fact, when $n - k - 1 > 50$, the value of c is so close to 2 that we can

use a simple *rule of thumb* for a 95% confidence interval: $\hat{\beta}_j$ plus or minus two of its standard errors. For small degrees of freedom, the exact percentiles should be obtained from the *t* tables.

It is easy to construct confidence intervals for any other level of confidence. For example, a 90% CI is obtained by choosing c to be the 95th percentile in the t_{n-k-1} distribution. When $df = n - k - 1 = 25$, $c = 1.71$, and so the 90% CI is $\hat{\beta}_j \pm 1.71 \cdot \text{se}(\hat{\beta}_j)$, which is necessarily narrower than the 95% CI. For a 99% CI, c is the 99.5th percentile in the t_{25} distribution. When $df = 25$, the 99% CI is roughly $\hat{\beta}_j \pm 2.79 \cdot \text{se}(\hat{\beta}_j)$, which is inevitably wider than the 95% CI.

Many modern regression packages save us from doing any calculations by reporting a 95% CI along with each coefficient and its standard error. Once a confidence interval is constructed, it is easy to carry out two-tailed hypotheses tests. If the null hypothesis is H_0: $\beta_j = a_j$, then H_0 is rejected against H_1: $\beta_j \neq a_j$ at (say) the 5% significance level if, and only if, a_j is *not* in the 95% confidence interval.

EXAMPLE 4.8
(Hedonic Price Model for Houses)

A model that explains the price of a good in terms of the good's characteristics is called an *hedonic price model*. The following equation is an hedonic price model for housing prices; the characteristics are square footage (*sqrft*), number of bedrooms (*bdrms*), and number of bathrooms (*bthrms*). Often *price* appears in logarithmic form, as do some of the explanatory variables. Using $n = 19$ observations on houses that were sold in Waltham, Massachusetts, in 1990, the estimated equation (with standard errors in parentheses below the coefficient estimates) is

$$\begin{gathered}\log(\hat{price}) = \underset{(1.15)}{7.46} + \underset{(.184)}{.634}\,\log(sqrft) - \underset{(.059)}{.066}\,bdrms + \underset{(.075)}{.158}\,bthrms \\ n = 19,\ R^2 = .806.\end{gathered}$$

Since *price* and *sqrft* both appear in logarithmic form, the price elasticity with respect to square footage is .634, so that, holding number of bedrooms and bathrooms fixed, a 1% increase in square footage increases the predicted housing price by about .634%. We can construct a 95% confidence interval for the population elasticity using the fact that the estimated model has $n - k - 1 = 19 - 3 - 1 = 15$ degrees of freedom. From Table G.2, we find the 97.5th percentile in the t_{15} distribution: $c = 2.131$. Thus, the 95% confidence interval for $\beta_{\log(sqrft)}$ is $.634 \pm 2.131(.184)$, or (.242,1.026). Since zero is excluded from this confidence interval, we reject H_0: $\beta_{\log(sqrft)} = 0$ against the two-sided alternative at the 5% level.

The coefficient on *bdrms* is negative, which seems counterintuitive. However, it is important to remember the ceteris paribus nature of this coefficient: it measures the effect of another bedroom, holding size of the house and number of bathrooms fixed. If two houses are the same size but one has more bedrooms, then the house with more bedrooms has smaller bedrooms; more bedrooms that are smaller is not necessarily a good thing. In any case, we can see that the 95% confidence interval for β_{bdrms} is fairly wide, and it contains the value zero: $-.066 \pm 2.131(.059)$ or (−.192,.060). Thus, *bdrms* does not have a statistically significant ceteris paribus effect on housing price.

Given size and number of bedrooms, one more bathroom is predicted to increase housing price by about 15.8%. (Remember that we must multiply the coefficient on *bthrms* by 100 to turn the effect into a percent.) The 95% confidence interval for β_{bthrms} is (−.002,.318). In this case, zero is barely in the confidence interval, so technically speaking $\hat{\beta}_{bthrms}$ is not statistically significant at the 5% level against a two-sided alternative. Since it is very close to being significant, we would probably conclude that number of bathrooms has an effect on log(*price*).

You should remember that a confidence interval is only as good as the underlying assumptions used to construct it. If we have omitted important factors that are correlated with the explanatory variables, then the coefficient estimates are not reliable: OLS is biased. If heteroskedasticity is present—for instance, in the previous example, if the variance of log(*price*) depends on any of the explanatory variables—then the standard error is not valid as an estimate of sd($\hat{\beta}_j$) (as we discussed in Section 3.4), and the confidence interval computed using these standard errors will not truly be a 95% CI. We have also used the normality assumption on the errors in obtaining these CIs, but, as we will see in Chapter 5, this is not as important for applications involving hundreds of observations.

4.4 TESTING HYPOTHESES ABOUT A SINGLE LINEAR COMBINATION OF THE PARAMETERS

The previous two sections have shown how to use classical hypothesis testing or confidence intervals to test hypotheses about a single β_j at a time. In applications, we must often test hypotheses involving more than one of the population parameters. In this section, we show how to test a single hypothesis involving more than one of the β_j. Section 4.5 shows how to test multiple hypotheses.

To illustrate the general approach, we will consider a simple model to compare the returns to education at junior colleges and four-year colleges; for simplicity, we refer to the latter as "universities." [This example is motivated by Kane and Rouse (1995), who provide a detailed analysis of this question.] The population includes working people with a high school degree, and the model is

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u, \tag{4.17}$$

where *jc* is number of years attending a two-year college and *univ* is number of years at a four-year college. Note that any combination of junior college and college is allowed, including $jc = 0$ and $univ = 0$.

The hypothesis of interest is whether a year at a junior college is worth a year at a university: this is stated as

$$H_0: \beta_1 = \beta_2. \tag{4.18}$$

Under H_0, another year at a junior college and another year at a university lead to the same ceteris paribus percentage increase in *wage*. For the most part, the alternative of

interest is one-sided: a year at a junior college is worth less than a year at a university. This is stated as

$$H_1: \beta_1 < \beta_2. \tag{4.19}$$

The hypotheses in (4.18) and (4.19) concern *two* parameters, β_1 and β_2, a situation we have not faced yet. We cannot simply use the individual t statistics for $\hat{\beta}_1$ and $\hat{\beta}_2$ to test H_0. However, conceptually, there is no difficulty in constructing a t statistic for testing (4.18). In order to do so, we rewrite the null and alternative as $H_0: \beta_1 - \beta_2 = 0$ and $H_1: \beta_1 - \beta_2 < 0$, respectively. The t statistic is based on whether the estimated difference $\hat{\beta}_1 - \hat{\beta}_2$ is sufficiently less than zero to warrant rejecting (4.18) in favor of (4.19). To account for the sampling error in our estimators, we standardize this difference by dividing by the standard error:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{se}(\hat{\beta}_1 - \hat{\beta}_2)}. \tag{4.20}$$

Once we have the t statistic in (4.20), testing proceeds as before. We choose a significance level for the test and, based on the *df*, obtain a critical value. Because the alternative is of the form in (4.19), the rejection rule is of the form $t < -c$, where c is a positive value chosen from the appropriate t distribution. Or, we compute the t statistic and then compute the p-value (see Section 4.2).

The only thing that makes testing the equality of two different parameters more difficult than testing about a single β_j is obtaining the standard error in the denominator of (4.20). Obtaining the numerator is trivial once we have peformed the OLS regression. For concreteness, suppose the following equation has been obtained using $n = 285$ individuals:

$$\begin{aligned} \widehat{\log(wage)} = \; & 1.43 + .098\, jc + .124\, univ + .019\, exper \\ & (0.27) \quad (.031) \quad (.035) \quad (.008) \\ & n = 285,\ R^2 = .243. \end{aligned} \tag{4.21}$$

It is clear from (4.21) that *jc* and *univ* have both economically and statistically significant effects on wage. This is certainly of interest, but we are more concerned about testing whether the estimated *difference* in the coefficients is statistically significant. The difference is estimated as $\hat{\beta}_1 - \hat{\beta}_2 = -.026$, so the return to a year at a junior college is about 2.6 percentage points less than a year at a university. Economically, this is not a trivial difference. The difference of $-.026$ is the numerator of the t statistic in (4.20).

Unfortunately, the regression results in equation (4.21) do *not* contain enough information to obtain the standard error of $\hat{\beta}_1 - \hat{\beta}_2$. It might be tempting to claim that $\text{se}(\hat{\beta}_1 - \hat{\beta}_2) = \text{se}(\hat{\beta}_1) - \text{se}(\hat{\beta}_2)$, but this does not make sense in the current example because $\text{se}(\hat{\beta}_1) - \text{se}(\hat{\beta}_2) = -.038$. Standard errors must *always* be positive because they are estimates of standard deviations. While the standard error of the difference $\hat{\beta}_1 - \hat{\beta}_2$ certainly depends on $\text{se}(\hat{\beta}_1)$ and $\text{se}(\hat{\beta}_2)$, it does so in a somewhat complicated way. To find $\text{se}(\hat{\beta}_1 - \hat{\beta}_2)$, we first obtain the variance of the difference. Using the results on variances in Appendix B, we have

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\,\text{Cov}(\hat{\beta}_1,\hat{\beta}_2). \tag{4.22}$$

Observe carefully how the two variances are *added* together, and twice the covariance is then subtracted. The standard deviation of $\hat{\beta}_1 - \hat{\beta}_2$ is just the square root of (4.22) and, since $[\text{se}(\hat{\beta}_1)]^2$ is an unbiased estimator of $\text{Var}(\hat{\beta}_1)$, and similarly for $[\text{se}(\hat{\beta}_2)]^2$, we have

$$\text{se}(\hat{\beta}_1 - \hat{\beta}_2) = \left\{[\text{se}(\hat{\beta}_1)]^2 + [\text{se}(\hat{\beta}_2)]^2 - 2s_{12}\right\}^{1/2}, \tag{4.23}$$

where s_{12} denotes an estimate of $\text{Cov}(\hat{\beta}_1,\hat{\beta}_2)$. We have not displayed a formula for $\text{Cov}(\hat{\beta}_1,\hat{\beta}_2)$. Some regression packages have features that allow one to obtain s_{12}, in which case one can compute the standard error in (4.23) and then the t statistic in (4.20). Appendix E shows how to use matrix algebra to obtain s_{12}.

We suggest another route that is much simpler to compute, less likely to lead to an error, and readily applied to a variety of problems. Rather than trying to compute $\text{se}(\hat{\beta}_1 - \hat{\beta}_2)$ from (4.23), it is much easier to estimate a different model that directly delivers the standard error of interest. Define a new parameter as the difference between β_1 and β_2: $\theta_1 = \beta_1 - \beta_2$. Then we want to test

$$\text{H}_0\text{: } \theta_1 = 0 \text{ against } \text{H}_1\text{: } \theta_1 < 0. \tag{4.24}$$

The t statistic (4.20) in terms of $\hat{\theta}_1$ is just $t = \hat{\theta}_1/\text{se}(\hat{\theta}_1)$. The challenge is finding $\text{se}(\hat{\theta}_1)$.

We can do this by rewriting the model so that θ_1 appears directly on one of the independent variables. Since $\theta_1 = \beta_1 - \beta_2$, we can also write $\beta_1 = \theta_1 + \beta_2$. Plugging this into (4.17) and rearranging gives the equation

$$\begin{aligned}\log(wage) &= \beta_0 + (\theta_1 + \beta_2)jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta_1 jc + \beta_2(jc + univ) + \beta_3 exper + u.\end{aligned} \tag{4.25}$$

The key insight is that the parameter we are interested in testing hypotheses about, θ_1, now multiplies the variable *jc*. The intercept is still β_0, and *exper* still shows up as being multiplied by β_3. More importantly, there is a new variable multiplying β_2, namely $jc + univ$. Thus, if we want to directly estimate θ_1 and obtain the standard error $\hat{\theta}_1$, then we must construct the new variable $jc + univ$ and include it in the regression model in place of *univ*. In this example, the new variable has a natural interpretation: it is *total* years of college, so define $totcoll = jc + univ$ and write (4.25) as

$$\log(wage) = \beta_0 + \theta_1 jc + \beta_2 totcoll + \beta_3 exper + u. \tag{4.26}$$

The parameter β_1 has disappeared from the model, while θ_1 appears explicitly. This model is really just a different way of writing the original model. The only reason we have defined this new model is that, when we estimate it, the coefficient on *jc* is $\hat{\theta}_1$ and, more importantly, $\text{se}(\hat{\theta}_1)$ is reported along with the estimate. The t statistic that we want is the one reported by any regression package on the variable *jc* (*not* the variable *totcoll*).

When we do this with the 285 observations used earlier, the result is

$$\begin{gathered}\log(\widehat{wage}) = 1.43 - .026\ jc + .124\ totcoll + .019\ exper \\ (0.27)\quad (.018)\quad\ (.035)\quad\ (.008) \\ n = 285,\ R^2 = .243.\end{gathered} \tag{4.27}$$

The only number in this equation that we could not get from (4.21) is the standard error for the estimate $-.026$, which is .018. The t statistic for testing (4.18) is $-.026/.018 = -1.44$. Against the one-sided alternative (4.19), the p-value is about .075, so there is some, but not strong, evidence against (4.18).

The intercept and slope estimate on *exper*, along with their standard errors, are the same as in (4.21). This fact *must* be true, and it provides one way of checking whether the transformed equation has been properly estimated. The coefficient on the new variable, *totcoll*, is the same as the coefficient on *univ* in (4.21), and the standard error is also the same. We know that this must happen by comparing (4.17) and (4.25).

It is quite simple to compute a 95% confidence interval for $\theta_1 = \beta_1 - \beta_2$. Using the standard normal approximation, the CI is obtained as usual: $\hat{\theta}_1 \pm 1.96\ \text{se}(\hat{\theta}_1)$, which in this case leads to $-.026 \pm .035$.

The strategy of rewriting the model so that it contains the parameter of interest works in all cases and is easy to implement. (See Problems 4.12 and 4.14 for other examples.)

4.5 TESTING MULTIPLE LINEAR RESTRICTIONS: THE *F* TEST

The t statistic associated with any OLS coefficient can be used to test whether the corresponding unknown parameter in the population is equal to any given constant (which is usually, but not always, zero). We have just shown how to test hypotheses about a single linear combination of the β_j by rearranging the equation and running a regression using transformed variables. But so far, we have only covered hypotheses involving a *single* restriction. Frequently, we wish to test *multiple* hypotheses about the underlying parameters $\beta_0, \beta_1, \ldots, \beta_k$. We begin with the leading case of testing whether a set of independent variables has no partial effect on a dependent variable.

Testing Exclusion Restrictions

We already know how to test whether a particular variable has no partial effect on the dependent variable: use the t statistic. Now we want to test whether a *group* of variables has no effect on the dependent variable. More precisely, the null hypothesis is that a set of variables has no effect on y, once another set of variables has been controlled.

As an illustration of why testing significance of a group of variables is useful, we consider the following model that explains major league baseball players' salaries:

$$\begin{aligned}\log(salary) = {} & \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \\ & \beta_4 hrunsyr + \beta_5 rbisyr + u,\end{aligned} \tag{4.28}$$

where *salary* is the 1993 total salary, *years* is years in the league, *gamesyr* is average games played per year, *bavg* is career batting average (for example, *bavg* = 250), *hrunsyr* is home runs per year, and *rbisyr* is runs batted in per year. Suppose we want to test the null hypothesis that, once years in the league and games per year have been controlled for, the statistics measuring performance—*bavg*, *hrunsyr*, and *rbisyr*—have no effect on salary. Essentially, the null hypothesis states that productivity as measured by baseball statistics has no effect on salary.

In terms of the parameters of the model, the null hypothesis is stated as

$$H_0: \beta_3 = 0, \beta_4 = 0, \beta_5 = 0. \tag{4.29}$$

The null (4.29) constitutes three **exclusion restrictions**: if (4.29) is true, then *bavg*, *hrunsyr*, and *rbisyr* have no effect on log(*salary*) after *years* and *gamesyr* have been controlled for and therefore should be excluded from the model. This is an example of a set of **multiple restrictions** because we are putting more than one restriction on the parameters in (4.28); we will see more general examples of multiple restrictions later. A test of multiple restrictions is called a **multiple hypotheses test** or a **joint hypotheses test**.

What should be the alternative to (4.29)? If what we have in mind is that "performance statistics matter, even after controlling for years in the league and games per year," then the appropriate alternative is simply

$$H_1: H_0 \text{ is not true.} \tag{4.30}$$

The alternative (4.30) holds if at least one of β_3, β_4, or β_5 is different from zero. (Any or all could be different from zero.) The test we study here is constructed to detect any violation of H_0. It is also valid when the alternative is something like H_1: $\beta_3 > 0$, or $\beta_4 > 0$, or $\beta_5 > 0$, but it will not be the best possible test under such alternatives. We do not have the space or statistical background necessary to cover tests that have more power under multiple one-sided alternatives.

How should we proceed in testing (4.29) against (4.30)? It is tempting to test (4.29) by using the *t* statistics on the variables *bavg*, *hrunsyr*, and *rbisyr* to determine whether each variable is *individually* significant. This option is not appropriate. A particular *t* statistic tests a hypothesis that puts no restrictions on the other parameters. Besides, we would have three outcomes to contend with—one for each *t* statistic. What would constitute rejection of (4.29) at, say, the 5% level? Should all three or only one of the three *t* statistics be required to be significant at the 5% level? These are hard questions, and fortunately we do not have to answer them. Furthermore, using separate *t* statistics to test a multiple hypothesis like (4.29) can be very misleading. We need a way to test the exclusion restrictions *jointly*.

To illustrate these issues, we estimate equation (4.28) using the data in MLB1.RAW. This gives

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.29)}{11.10} + \underset{(.0121)}{.0689}\, years + \underset{(.0026)}{.0126}\, gamesyr \\ &+ \underset{(.00110)}{.00098}\, bavg + \underset{(.0161)}{.0144}\, hrunsyr + \underset{(.0072)}{.0108}\, rbisyr \\ n &= 353,\ \text{SSR} = 183.186,\ R^2 = .6278, \end{aligned} \tag{4.31}$$

where SSR is the sum of squared residuals. (We will use this later.) We have left several terms after the decimal in SSR and *R*-squared to facilitate future comparisons. Equation (4.31) reveals that, while *years* and *gamesyr* are statistically significant, none of the variables *bavg*, *hrunsyr*, and *rbisyr* has a statistically significant *t* statistic against a two-sided alternative, at the 5% significance level. (The *t* statistic on *rbisyr* is the closest to being significant; its two-sided *p*-value is .134.) Thus, based on the three *t* statistics, it appears that we cannot reject H_0.

This conclusion turns out to be wrong. In order to see this, we must derive a test of multiple restrictions whose distribution is known and tabulated. The sum of squared residuals now turns out to provide a very convenient basis for testing multiple hypotheses. We will also show how the *R*-squared can be used in the special case of testing for exclusion restrictions.

Knowing the sum of squared residuals in (4.31) tells us nothing about the truth of the hypothesis in (4.29). However, the factor that will tell us something is how much the SSR increases when we drop the variables *bavg*, *hrunsyr*, and *rbisyr* from the model. Remember that, because the OLS estimates are chosen to minimize the sum of squared residuals, the SSR *always* increases when variables are dropped from the model; this is an algebraic fact. The question is whether this increase is large enough, *relative* to the SSR in the model with all of the variables, to warrant rejecting the null hypothesis.

The model without the three variables in question is simply

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + u. \tag{4.32}$$

In the context of hypothesis testing, equation (4.32) is the **restricted model** for testing (4.29); model (4.28) is called the **unrestricted model**. The restricted model always has fewer parameters than the unrestricted model.

When we estimate the restricted model using the data in MLB1.RAW, we obtain

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.11)}{11.22} + \underset{(.0125)}{.0713}\, years + \underset{(.0013)}{.0202}\, gamesyr \\ n &= 353,\ \text{SSR} = 198.311,\ R^2 = .5971. \end{aligned} \tag{4.33}$$

As we surmised, the SSR from (4.33) is greater than the SSR from (4.31), and the *R*-squared from the restricted model is less than the *R*-squared from the unrestricted model. What we need to decide is whether the increase in the SSR in going from the unrestricted model to the restricted model (183.186 to 198.311) is large enough to warrant rejection of (4.29). As with all testing, the answer depends on the significance level of the test. But we cannot carry out the test at a chosen significance level until we have a statistic whose distribution is known, and can be tabulated, under H_0. Thus, we need a way to combine the information in the two SSRs to obtain a test statistic with a known distribution under H_0.

Since it is no more difficult, we might as well derive the test for the general case. Write the *unrestricted* model with *k* independent variables as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u; \tag{4.34}$$

the number of parameters in the unrestricted model is $k + 1$. (Remember to add one for the intercept.) Suppose that we have q exclusion restrictions to test: that is, the null hypothesis states that q of the variables in (4.34) have zero coefficients. For notational simplicity, assume that it is the last q variables in the list of independent variables: $x_{k-q+1}, \ldots, x_k$. (The order of the variables, of course, is arbitrary and unimportant.) The null hypothesis is stated as

$$H_0: \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \tag{4.35}$$

which puts q exclusion restrictions on the model (4.34). The alternative to (4.35) is simply that it is false; this means that at least one of the parameters listed in (4.35) is different from zero. When we impose the restrictions under H_0, we are left with the restricted model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + u. \tag{4.36}$$

In this subsection, we assume that both the unrestricted and restricted models contain an intercept, since that is the case most widely encountered in practice.

Now for the test statistic itself. Earlier, we suggested that looking at the relative increase in the SSR when moving from the unrestricted to the restricted model should be informative for testing the hypothesis (4.35). The ***F* statistic** (or *F ratio*) is defined by

$$F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)}, \tag{4.37}$$

where SSR_r is the sum of squared residuals from the restricted model and SSR_{ur} is the sum of squared residuals from the unrestricted model.

QUESTION 4.4

Consider relating individual performance on a standardized test, *score*, to a variety of other variables. School factors include average class size, per student expenditures, average teacher compensation, and total school enrollment. Other variables specific to the student are family income, mother's education, father's education, and number of siblings. The model is

$$score = \beta_0 + \beta_1 classize + \beta_2 expend + \beta_3 tchcomp + \beta_4 enroll + \beta_5 faminc + \beta_6 motheduc + \beta_7 fatheduc + \beta_8 siblings + u.$$

State the null hypothesis that student-specific variables have no effect on standardized test performance, once school-related factors have been controlled for. What are k and q for this example? Write down the restricted version of the model.

You should immediately notice that, since SSR_r can be no smaller than SSR_{ur}, the F statistic is *always* nonnegative (and almost always strictly positive). Thus, if you compute a negative F statistic, then something is wrong; the order of the SSRs in the numerator of F has usually been reversed. Also, the SSR in the denominator of F is the SSR from the *unrestricted* model. The easiest way to remember where the SSRs appear is to think of F as measuring the relative increase in SSR when moving from the unrestricted to the restricted model.

The difference in SSRs in the numerator of F is divided by q, which is the number of restrictions imposed in moving from the unrestricted to the restricted model (q independent variables are dropped). Therefore, we can write

$$q = \textbf{numerator degrees of freedom} = df_r - df_{ur}, \tag{4.38}$$

which also shows that q is the difference in degrees of freedom between the restricted and unrestricted models. (Recall that df = number of observations − number of estimated parameters.) Since the restricted model has fewer parameters—and each model is estimated using the same n observations—df_r is always greater than df_{ur}.

The SSR in the denominator of F is divided by the degrees of freedom in the unrestricted model:

$$n - k - 1 = \textbf{denominator degrees of freedom} = df_{ur}. \tag{4.39}$$

In fact, the denominator of F is just the unbiased estimator of $\sigma^2 = \text{Var}(u)$ in the unrestricted model.

In a particular application, computing the F statistic is easier than wading through the somewhat cumbersome notation used to describe the general case. We first obtain the degrees of freedom in the unrestricted model, df_{ur}. Then, we count how many variables are excluded in the restricted model; this is q. The SSRs are reported with every OLS regression, and so forming the F statistic is simple.

In the major league baseball salary regression, $n = 353$, and the full model (4.28) contains six parameters. Thus, $n - k - 1 = df_{ur} = 353 - 6 = 347$. The restricted model (4.32) contains three fewer independent variables than (4.28), and so $q = 3$. Thus, we have all of the ingredients to compute the F statistic; we hold off doing so until we know what to do with it.

In order to use the F statistic, we must know its sampling distribution under the null in order to choose critical values and rejection rules. It can be shown that, under H_0 (and assuming the CLM assumptions hold), F is distributed as an F random variable with $(q, n - k - 1)$ degrees of freedom. We write this as

$$F \sim F_{q,n-k-1}.$$

The distribution of $F_{q,n-k-1}$ is readily tabulated and available in statistical tables (see Table G.3) and, even more importantly, in statistical software.

We will not derive the F distribution because the mathematics is very involved. Basically, it can be shown that equation (4.37) is actually the ratio of two independent chi-square random variables, divided by their respective degrees of freedom. The numerator chi-square random variable has q degrees of freedom, and the chi-square in the denominator has $n - k - 1$ degrees of freedom. This is the definition of an F distributed random variable (see Appendix B).

It is pretty clear from the definition of F that we will reject H_0 in favor of H_1 when F is sufficiently "large." How large depends on our chosen significance level. Suppose that we have decided on a 5% level test. Let c be the 95$^{\text{th}}$ percentile in the $F_{q,n-k-1}$ distribution. This critical value depends on q (the numerator df) and $n - k - 1$ (the denominator df). It is important to keep the numerator and denominator degrees of freedom straight.

The 10%, 5%, and 1% critical values for the F distribution are given in Table G.3. The rejection rule is simple. Once c has been obtained, we reject H_0 in favor of H_1 at the chosen significance level if

$$F > c. \tag{4.40}$$

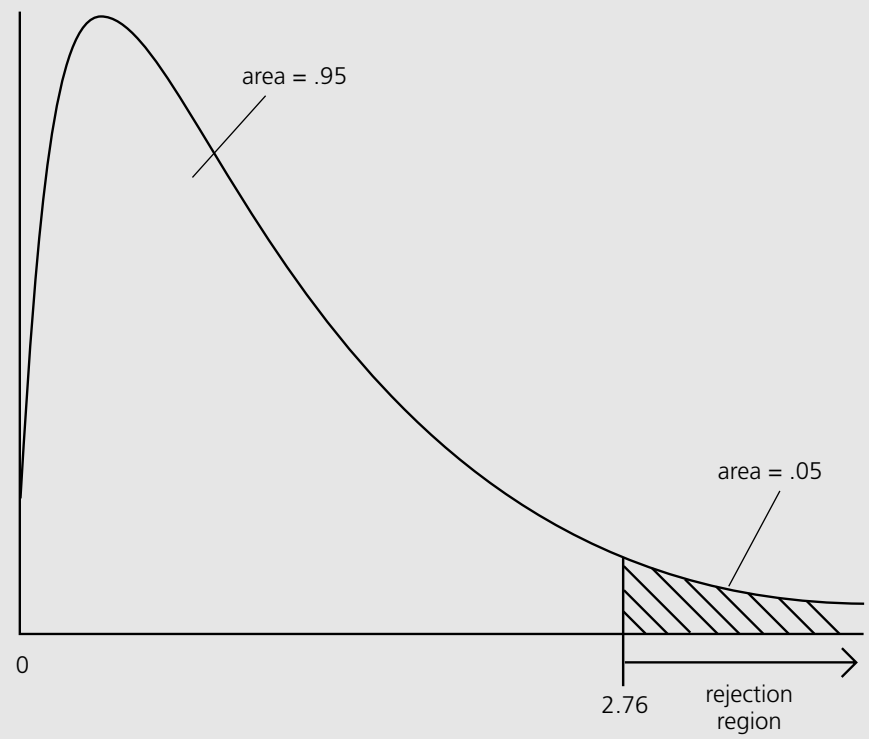
With a 5% significance level, $q = 3$, $n - k - 1 = 60$, and the critical value is $c = 2.76$. We would reject H_0 at the 5% level if the computed value of the F statistic exceeds 2.76. The 5% critical value and rejection region are shown in Figure 4.7. For the same degrees of freedom, the 1% critical value is 4.13.

In most applications, the numerator degrees of freedom (q) will be notably smaller than the denominator degrees of freedom ($n - k - 1$). Applications where $n - k - 1$ is small are unlikely to be successful because the parameters in the null model will probably not be precisely estimated. When the denominator *df* reaches about 120, the F distribution is no longer sensitive to it. (This is entirely analogous to the t distribution being well-approximated by the standard normal distribution as the *df* gets large.) Thus, there is an entry in the table for the denominator $df = \infty$, and this is what we use with large samples (since $n - k - 1$ is then large). A similar statement holds for a very large numerator *df*, but this rarely occurs in applications.

If H_0 is rejected, then we say that $x_{k-q+1}, \ldots, x_k$ are **jointly statistically significant** (or just *jointly significant)* at the appropriate significance level. This test alone does not

Figure 4.7

The 5% critical value and rejection region in an $F_{3,60}$ distribution.



allow us to say which of the variables has a partial effect on y; they may all affect y or maybe only one affects y. If the null is not rejected, then the variables are **jointly insignificant**, which often justifies dropping them from the model.

For the major league baseball example with three numerator degrees of freedom and 347 denominator degrees of freedom, the 5% critical value is 2.60, and the 1% critical value is 3.78. We reject H_0 at the 1% level if F is above 3.78; we reject at the 5% level if F is above 2.60.

We are now in a position to test the hypothesis that we began this section with: after controlling for *years* and *gamesyr*, the variables *bavg*, *hrunsyr*, and *rbisyr* have no effect on players' salaries. In practice, it is easiest to first compute $(SSR_r - SSR_{ur})/SSR_{ur}$ and to multiply the result by $(n - k - 1)/q$; the reason the formula is stated as in (4.37) is that it makes it easier to keep the numerator and denominator degrees of freedom straight. Using the SSRs in (4.31) and (4.33), we have

$$F = \frac{(198.311 - 183.186)}{183.186} \cdot \frac{347}{3} \approx 9.55.$$

This number is well above the 1% critical value in the F distribution with 3 and 347 degrees of freedom, and so we soundly reject the hypothesis that *bavg*, *hrunsyr*, and *rbisyr* have no effect on salary.

The outcome of the joint test may seem surprising in light of the insignificant t statistics for the three variables. What is happening is that the two variables *hrunsyr* and *rbisyr* are highly correlated, and this multicollinearity makes it difficult to uncover the partial effect of each variable; this is reflected in the individual t statistics. The F statistic tests whether these variables (including *bavg*) are *jointly* significant, and multicollinearity between *hrunsyr* and *rbisyr* is much less relevant for testing this hypothesis. In Problem 4.16, you are asked to reestimate the model while dropping *rbisyr*, in which case *hrunsyr* becomes very significant. The same is true for *rbisyr* when *hrunsyr* is dropped from the model.

The F statistic is often useful for testing exclusion of a group of variables when the variables in the group are highly correlated. For example, suppose we want to test whether firm performance affects the salaries of chief executive officers. There are many ways to measure firm performance, and it probably would not be clear ahead of time which measures would be most important. Since measures of firm performance are likely to be highly correlated, hoping to find individually significant measures might be asking too much due to multicollinearity. But an F test can be used to determine whether, as a group, the firm performance variables affect salary.

Relationship Between *F* and *t* Statistics

We have seen in this section how the F statistic can be used to test whether a group of variables should be included in a model. What happens if we apply the F statistic to the case of testing significance of a *single* independent variable? This case is certainly not ruled out by the previous development. For example, we can take the null to be H_0: $\beta_k = 0$ and $q = 1$ (to test the single exclusion restriction that x_k can be excluded from the model). From Section 4.2, we know that the t statistic on β_k can be used to test this hypothesis. The question, then, is do we have two separate ways of testing hypotheses

about a single coefficient? The answer is no. It can be shown that the F statistic for testing exclusion of a single variable is equal to the *square* of the corresponding t statistic. Since t^2_{n-k-1} has an $F_{1,n-k-1}$ distribution, the two approaches lead to exactly the same outcome, provided that the alternative is two-sided. The t statistic is more flexible for testing a single hypothesis because it can be used to test against one-sided alternatives. Since t statistics are also easier to obtain than F statistics, there is really no reason to use an F statistic to test hypotheses about a single parameter.

The *R*-Squared Form of the *F* Statistic

In most applications, it turns out to be more convenient to use a form of the F statistic that can be computed using the R-squareds from the restricted and unrestricted models. One reason for this is that the R-squared is always between zero and one, whereas the SSRs can be very large depending on the units of measurement of y, making the calculation based on the SSRs tedious. Using the fact that $SSR_r = SST(1 - R_r^2)$ and $SSR_{ur} = SST(1 - R_{ur}^2)$, we can substitute into (4.37) to obtain

$$F \equiv \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)} \tag{4.41}$$

(note that the SST terms cancel everywhere). This is called the ***R*-squared form of the *F* statistic**.

Since the R-squared is reported with almost all regressions (whereas the SSR is not), it is easy to use the R-squareds from the unrestricted and restricted models to test for exclusion of some variables. Particular attention should be paid to the order of the R-squareds in the numerator: the *unrestricted* R-squared comes first [contrast this with the SSRs in (4.37)]. Since $R_{ur}^2 > R_r^2$, this shows again that F will always be positive.

In using the R-squared form of the test for excluding a set of variables, it is important to *not* square the R-squared before plugging it into formula (4.41); the squaring has already been done. All regressions report R^2, and these numbers are plugged directly into (4.41). For the baseball salary example, we can use (4.41) to obtain the F statistic:

$$F = \frac{(.6278 - .5971)}{1 - .6278} \cdot \frac{347}{3} \approx 9.54,$$

which is very close to what we obtained before. (The difference is due to a rounding error.)

EXAMPLE 4.9
(Parents' Education in a Birth Weight Equation)

As another example of computing an F statistic, consider the following model to explain child birth weight in terms of various factors:

$$bwght = \beta_0 + \beta_1 cigs + \beta_2 parity + \beta_3 faminc + \beta_4 motheduc + \beta_5 fatheduc + u, \tag{4.42}$$

where *bwght* is birth weight, in pounds, *cigs* is average number of cigarettes the mother smoked per day during pregnancy, *parity* is the birth order of this child, *faminc* is annual family income, *motheduc* is years of schooling for the mother, and *fatheduc* is years of schooling for the father. Let us test the null hypothesis that, after controlling for *cigs*, *parity*, and *faminc*, parents' education has no effect on birth weight. This is stated as H_0: $\beta_4 = 0$, $\beta_5 = 0$, and so there are $q = 2$ exclusion restrictions to be tested. There are $k + 1 = 6$ parameters in the unrestricted model (4.42), so the *df* in the unrestricted model is $n - 6$, where n is the sample size.

We will test this hypothesis using the data in BWGHT.RAW. This data set contains information on 1,388 births, but we must be careful in counting the observations used in testing the null hypothesis. It turns out that information on at least one of the variables *motheduc* and *fatheduc* is missing for 197 births in the sample; these observations cannot be included when estimating the unrestricted model. Thus, we really have $n = 1{,}191$ observations, and so there are $1{,}191 - 6 = 1{,}185$ *df* in the unrestricted model. We must be sure to use these *same* 1,191 observations when estimating the restricted model (not the full 1,388 observations that are available). Generally, when estimating the restricted model to compute an *F* test, we must use the same observations to estimate the unrestricted model; otherwise the test is not valid. When there are no missing data, this will not be an issue.

The numerator *df* is 2, and the denominator *df* is 1,185; from Table G.3, the 5% critical value is $c = 3.0$. Rather than report the complete results, for brevity we present only the *R*-squareds. The *R*-squared for the full model turns out to be $R^2_{ur} = .0387$. When *motheduc* and *fatheduc* are dropped from the regression, the *R*-squared falls to $R^2_r = .0364$. Thus, the *F* statistic is $F = [(.0387 - .0364)/(1 - .0387)](1{,}185/2) = 1.42$; since this is well below the 5% critical value, we fail to reject H_0. In other words, *motheduc and fatheduc* are jointly insignificant in the birth weight equation.

Computing *p*-values for *F* Tests

For reporting the outcomes of *F* tests, *p*-values are especially useful. Since the *F* distribution depends on the numerator and denominator *df*, it is difficult to get a feel for how strong or weak the evidence is against the null hypothesis simply by looking at the value of the *F* statistic and one or two critical values.

In the *F* testing context, the *p*-value is defined as

$$p\text{-value} = P(\mathcal{F} > F), \quad \textbf{(4.43)}$$

where, for emphasis, we let $\mathcal{F}$ denote an *F* random variable with $(q, n - k - 1)$ degrees of freedom, and *F* is the actual value of the test statistic. The *p*-value still has the same interpretation as it did for *t* statistics: it is the probability of observing

QUESTION 4.5

The data in ATTEND.RAW were used to estimate the two equations

$$\widehat{atndrte} = \underset{(2.87)}{47.13} + \underset{(1.09)}{13.37}\, priGPA$$
$$n = 680,\ R^2 = .183,$$

and

$$\widehat{atndrte} = \underset{(3.88)}{75.70} + \underset{(1.08)}{17.26}\, priGPA - \underset{(?)}{1.72}\, ACT,$$
$$n = 680,\ R^2 = .291,$$

where, as always, standard errors are in parentheses; the standard error for *ACT* is missing in the second equation. What is the *t* statistic for the coefficient on *ACT*? (Hint: First compute the *F* statistic for significance of *ACT*.)

a value of the F at least as large as we did, *given* that the null hypothesis is true. A small p-value is evidence against H_0. For example, p-value = .016 means that the chance of observing a value of F as large as we did when the null hypothesis was true is only 1.6%; we usually reject H_0 in such cases. If the p-value = .314, then the chance of observing a value of the F statistic as large as we did under the null hypothesis is 31.4%. Most would find this to be pretty weak evidence against H_0.

As with t testing, once the p-value has been computed, the F test can be carried out at any significance level. For example, if the p-value = .024, we reject H_0 at the 5% significance level but not at the 1% level.

The p-value for the F test in Example 4.9 is .238, and so the null hypothesis that $\beta_{motheduc}$ and $\beta_{fatheduc}$ are both zero is not rejected at even the 20% significance level.

Many econometrics packages have a built-in feature for testing multiple exclusion restrictions. These packages have several advantages over calculating the statistics by hand: we will less likely make a mistake, p-values are computed automatically, and the problem of missing data, as in Example 4.9, is handled without any additional work on our part.

The *F* Statistic for Overall Significance of a Regression

A special set of exclusion restrictions is routinely tested by most regression packages. These restrictions have the same interpretation, regardless of the model. In the model with k independent variables, we can write the null hypothesis as

$$H_0: x_1, x_2, \ldots, x_k \text{ do not help to explain } y.$$

This null hypothesis is, in a way, very pessimistic. It states that *none* of the explanatory variables has an effect on y. Stated in terms of the parameters, the null is that all slope parameters are zero:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0, \tag{4.44}$$

and the alternative is that at least one of the β_j is different from zero. Another useful way of stating the null is that $H_0: E(y|x_1,x_2, \ldots, x_k) = E(y)$, so that knowing the values of x_1, $x_2, \ldots, x_k$ does not affect the expected value of y.

There are k restrictions in (4.44), and when we impose them, we get the restricted model

$$y = \beta_0 + u; \tag{4.45}$$

all independent variables have been dropped from the equation. Now, the R-squared from estimating (4.45) is zero; none of the variation in y is being explained because there are no explanatory variables. Therefore, the F statistic for testing (4.44) can be written as

$$\frac{R^2/k}{(1 - R^2)/(n - k - 1)}, \tag{4.46}$$

where R^2 is just the usual R-squared from the regression of y on $x_1, x_2, \ldots, x_k$.

Most regression packages report the F statistic in (4.46) automatically, which makes it tempting to use this statistic to test general exclusion restrictions. You must avoid this temptation. The F statistic in (4.41) is used for general exclusion restrictions; it depends on the R-squareds from the restricted and unrestricted models. The special form of (4.46) is valid only for testing joint exclusion of *all* independent variables. This is sometimes called testing the **overall significance of the regression**.

If we fail to reject (4.44), then there is no evidence that any of the independent variables help to explain y. This usually means that we must look for other variables to explain y. For Example 4.9, the F statistic for testing (4.44) is about 9.55 with $k = 5$ and $n - k - 1 = 1{,}185$ *df*. The p-value is zero to four places after the decimal point, so that (4.44) is rejected very strongly. Thus, we conclude that the variables in the *bwght* equation *do* explain some variation in *bwght*. The amount explained is not large: only 3.87%. But the seemingly small R-squared results in a highly significant F statistic. That is why we must compute the F statistic to test for joint significance and not just look at the size of the R-squared.

Occasionally, the F statistic for the hypothesis that all independent variables are jointly insignificant is the focus of a study. Problem 4.10 asks you to use stock return data to test whether stock returns over a four-year horizon are predictable based on information known only at the beginning of the period. Under the *efficient markets hypothesis*, the returns should not be predictable; the null hypothesis is precisely (4.44).

Testing General Linear Restrictions

Testing exclusion restrictions is by far the most important application of F statistics. Sometimes, however, the restrictions implied by a theory are more complicated than just excluding some independent variables. It is still straightforward to use the F statistic for testing.

As an example, consider the following equation:

$$\log(price) = \beta_0 + \beta_1 \log(assess) + \beta_2 \log(lotsize) + \beta_3 \log(sqrft) + \beta_4 bdrms + u, \tag{4.47}$$

where *price* is house price, *assess* is the assessed housing value (before the house was sold), *lotsize* is size of the lot, in feet, *sqrft* is square footage, and *bdrms* is number of bedrooms. Now, suppose we would like to test whether the assessed housing price is a rational valuation. If this is the case, then a 1% change in *assess* should be associated with a 1% change in *price*; that is, $\beta_1 = 1$. In addition, *lotsize*, *sqrft*, and *bdrms* should not help to explain log(*price*), once the assessed value has been controlled for. Together, these hypotheses can be stated as

$$H_0: \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0. \tag{4.48}$$

There are four restrictions here to be tested; three are exclusion restrictions, but $\beta_1 = 1$ is not. How can we test this hypothesis using the F statistic?

As in the exclusion restriction case, we estimate the unrestricted model, (4.47) in this case, and then impose the restrictions in (4.48) to obtain the restricted model. It is

the second step that can be a little tricky. But all we do is plug in the restrictions. If we write (4.47) as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u, \tag{4.49}$$

then the restricted model is $y = \beta_0 + x_1 + u$. Now, in order to impose the restriction that the coefficient on x_1 is unity, we must estimate the following model:

$$y - x_1 = \beta_0 + u. \tag{4.50}$$

This is just a model with an intercept (β_0) but with a different dependent variable than in (4.49). The procedure for computing the F statistic is the same: estimate (4.50), obtain the SSR (SSR_r), and use this with the unrestricted SSR from (4.49) in the F statistic (4.37). We are testing $q = 4$ restrictions, and there are $n - 5$ *df* in the unrestricted model. The F statistic is simply $[(SSR_r - SSR_{ur})/SSR_{ur}][(n - 5)/4]$.

Before illustrating this test using a data set, we must emphasize one point: we cannot use the R-squared form of the F statistic for this example because the dependent variable in (4.50) is different from the one in (4.49). This means the total sum of squares from the two regressions will be different, and (4.41) is no longer equivalent to (4.37). As a general rule, the SSR form of the F statistic should be used if a different dependent variable is needed in running the restricted regression.

The estimated unrestricted model using the data in HPRICE1.RAW is

$$\begin{aligned} \widehat{\log(price)} = &- .034 + 1.043\ \log(assess) + .0074\ \log(lotsize) \\ &(.972) \quad (.151) \quad (.0386) \\ &- .1032\ \log(sqrft) + .0338\ bdrms \\ &(.1384) \quad (.0221) \\ n = 88,\ &\text{SSR} = 1.822,\ R^2 = .773. \end{aligned}$$

If we use separate t statistics to test each hypothesis in (4.48), we fail to reject each one. But rationality of the assessment is a joint hypothesis, so we should test the restrictions jointly. The SSR from the restricted model turns out to be $SSR_r = 1.880$, and so the F statistic is $[(1.880 - 1.822)/1.822](83/4) = .661$. The 5% critical value in an F distribution with (4,83) *df* is about 2.50, and so we fail to reject H_0. There is essentially no evidence against the hypothesis that the assessed values are rational.

4.6 REPORTING REGRESSION RESULTS

We end this chapter by providing a few guidelines on how to report multiple regression results for relatively complicated empirical projects. This should teach you to read published works in the applied social sciences, while also preparing you to write your own empirical papers. We will expand on this topic in the remainder of the text by reporting results from various examples, but many of the key points can be made now.

Naturally, the estimated OLS coefficients should always be reported. For the key variables in an analysis, you should *interpret* the estimated coefficients (which often requires knowing the units of measurement of the variables). For example, is an esti-

mate an elasticity, or does it have some other interpretation that needs explanation? The economic or practical importance of the estimates of the key variables should be discussed.

The standard errors should always be included along with the estimated coefficients. Some authors prefer to report the t statistics rather than the standard errors (and often just the absolute value of the t statistics). While nothing is really wrong with this, there is some preference for reporting standard errors. First, it forces us to think carefully about the null hypothesis being tested; the null is not always that the population parameter is zero. Second, having standard errors makes it easier to compute confidence intervals.

The R-squared from the regression should always be included. We have seen that, in addition to providing a goodness-of-fit measure, it makes calculation of F statistics for exclusion restrictions simple. Reporting the sum of squared residuals and the standard error of the regression is sometimes a good idea, but it is not crucial. The number of observations used in estimating any equation should appear near the estimated equation.

If only a couple of models are being estimated, the results can be summarized in equation form, as we have done up to this point. However, in many papers, several equations are estimated with many different sets of independent variables. We may estimate the same equation for different groups of people, or even have equations explaining different dependent variables. In such cases, it is better to summarize the results in one or more tables. The dependent variable should be indicated clearly in the table, and the independent variables should be listed in the first column. Standard errors (or t statistics) can be put in parentheses below the estimates.

EXAMPLE 4.10
(Salary-Pension Tradeoff for Teachers)

Let *totcomp* denote average total annual compensation for a teacher, including salary and all fringe benefits (pension, health insurance, and so on). Extending the standard wage equation, total compensation should be a function of productivity and perhaps other characteristics. As is standard, we use logarithmic form:

$$\log(totcomp) = f(productivity\ characteristics, other\ factors),$$

where $f(\cdot)$ is some function (unspecified for now). Write

$$totcomp = salary + benefits = salary\left(1 + \frac{benefits}{salary}\right).$$

This equation shows that total compensation is the product of two terms: *salary* and 1 + *b/s*, where *b/s* is shorthand for the "benefits to salary ratio." Taking the log of this equation gives log(*totcomp*) = log(*salary*) + log(1 + *b/s*). Now, for "small" *b/s*, log(1 + *b/s*) $\approx$ *b/s*; we will use this approximation. This leads to the econometric model

$$\log(salary) = \beta_0 + \beta_1(b/s) + other\ factors.$$

Testing the wage-benefits tradeoff then is the same as a test of H_0: $\beta_1 = -1$ against H_1: $\beta_1 \neq -1$.

We use the data in MEAP93.RAW to test this hypothesis. These data are averaged at the school level, and we do not observe very many other factors that could affect total compensation. We will include controls for size of the school (*enroll*), staff per thousand students (*staff*), and measures such as the school dropout and graduation rates. The average *b/s* in the sample is about .205, and the largest value is .450.

The estimated equations are given in Table 4.1, where standard errors are given in parentheses below the coefficient estimates. The key variable is *b/s*, the benefits-salary ratio.

From the first column in Table 4.1, we see that, without controlling for any other factors, the OLS coefficient for *b/s* is −.825. The *t* statistic for testing the null hypothesis H_0: $\beta_1 = -1$ is $t = (-.825 + 1)/.200 \approx .875$, and so the simple regression fails to reject H_0. After adding controls for school size and staff size (which roughly captures the number of students taught by each teacher), the estimate of the *b/s* coef-

Table 4.1

Testing the Salary-Benefits Tradeoff

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
b/s	−.825 (.200)	−.605 (.165)	−.589 (.165)
log(*enroll*)	———	.0874 (.0073)	.0881 (.0073)
log(*staff*)	———	−.222 (.050)	−.218 (.050)
droprate	———	———	−.00028 (.00161)
gradrate	———	———	.00097 (.00066)
intercept	10.523 (0.042)	10.884 (0.252)	10.738 (0.258)
Observations *R*-Squared	408 .040	408 .353	408 .361

QUESTION 4.6

How does adding *droprate* and *gradrate* affect the estimate of the salary-benefits tradeoff? Are these variables jointly significant at the 5% level? What about the 10% level?

ficient becomes −.605. Now the test of $\beta_1 = -1$ gives a t statistic of about 2.39; thus, H_0 is rejected at the 5% level against a two-sided alternative. The variables log(*enroll*) and log(*staff*) are very statistically significant.

SUMMARY

In this chapter, we have covered the very important topic of statistical inference, which allows us to infer something about the population model from a random sample. We summarize the main points:

1. Under the classical linear model assumptions MLR.1 through MLR.6, the OLS estimators are normally distributed.

2. Under the CLM assumptions, the t statistics have t distributions under the null hypothesis.

3. We use t statistics to test hypotheses about a single parameter against one- or two-sided alternatives, using one- or two-tailed tests, respectively. The most common null hypothesis is H_0: $\beta_j = 0$, but we sometimes want to test other values of β_j under H_0.

4. In classical hypothesis testing, we first choose a significance level, which, along with the *df* and alternative hypothesis, determines the critical value against which we compare the t statistic. It is more informative to compute the p-value for a t test—the smallest significance level for which the null hypothesis is rejected—so that the hypothesis can be tested at any significance level.

5. Under the CLM assumptions, confidence intervals can be constructed for each β_j. These CIs can be used to test any null hypothesis concerning β_j against a two-sided alternative.

6. Single hypothesis tests concerning more than one β_j can always be tested by rewriting the model to contain the parameter of interest. Then, a standard t statistic can be used.

7. The F statistic is used to test multiple exclusion restrictions, and there are two equivalent forms of the test. One is based on the SSRs from the restricted and unrestricted models. A more convenient form is based on the R-squareds from the two models.

8. When computing an F statistic, the numerator *df* is the number of restrictions being tested, while the denominator *df* is the degrees of freedom in the unrestricted model.

9. The alternative for F testing is two-sided. In the classical approach, we specify a significance level which, along with the numerator *df* and the denominator *df*, determines the critical value. The null hypothesis is rejected when the statistic, F, exceeds the critical value, c. Alternatively, we can compute a p-value to summarize the evidence against H_0.

10. General multiple linear restrictions can be tested using the sum of squared residuals form of the F statistic.

11. The *F* statistic for the overall significance of a regression tests the null hypothesis that *all* slope parameters are zero, with the intercept unrestricted. Under H_0, the explanatory variables have no effect on the expected value of *y*.

KEY TERMS

Alternative Hypothesis
Classical Linear Model
Classical Linear Model (CLM) Assumptions
Confidence Interval (CI)
Critical Value
Denominator Degrees of Freedom
Economic Significance
Exclusion Restrictions
F Statistic
Joint Hypotheses Test
Jointly Insignificant
Jointly Statistically Significant
Minimum Variance Unbiased Estimators
Multiple Hypotheses Test
Multiple Restrictions
Normality Assumption
Null Hypothesis
Numerator Degrees of Freedom
One-Sided Alternative
One-Tailed Test
Overall Significance of the Regression
p-Value
Practical Significance
R-squared Form of the *F* Statistic
Rejection Rule
Restricted Model
Significance Level
Statistically Insignificant
Statistically Significant
t Ratio
t Statistic
Two-Sided Alternative
Two-Tailed Test
Unrestricted Model

PROBLEMS

4.1 Which of the following can cause the usual OLS *t* statistics to be invalid (that is, not to have *t* distributions under H_0)?

(i) Heteroskedasticity.
(ii) A sample correlation coefficient of .95 between two independent variables that are in the model.
(iii) Omitting an important explanatory variable.

4.2 Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (*roe*, in percent form), and return on the firm's stock (*ros*, in percent form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

(i) In terms of the model parameters, state the null hypothesis that, after controlling for *sales* and *roe*, *ros* has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.
(ii) Using the data in CEOSAL1.RAW, the following equation was obtained by OLS:

$$\widehat{\log(salary)} = \underset{(0.32)}{4.32} + \underset{(.035)}{.280}\log(sales) + \underset{(.0041)}{.0174}\,roe + \underset{(.00054)}{.00024}\,ros$$
$$n = 209,\ R^2 = .283$$

By what percent is *salary* predicted to increase, if *ros* increases by 50 points? Does *ros* have a practically large effect on *salary*?

(iii) Test the null hypothesis that *ros* has no effect on *salary*, against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level.

(iv) Would you include *ros* in a final model explaining CEO compensation in terms of firm performance? Explain.

4.3 The variable *rdintens* is expenditures on research and development (R&D) as a percentage of sales. Sales are measured in millions of dollars. The variable *profmarg* is profits as a percentage of sales.

Using the data in RDCHEM.RAW for 32 firms in the chemical industry, the following equation is estimated:

$$\widehat{rdintens} = \underset{(1.369)}{.472} + \underset{(.216)}{.321}\log(sales) + \underset{(.046)}{.050}\,profmarg$$
$$n = 32,\ R^2 = .099$$

(i) Interpret the coefficient on log(*sales*). In particular, if *sales* increases by 10%, what is the estimated percentage point change in *rdintens*? Is this an economically large effect?

(ii) Test the hypothesis that R&D intensity does not change with *sales*, against the alternative that it does increase with sales. Do the test at the 5% and 10% levels.

(iii) Does *profmarg* have a statistically significant effect on *rdintens*?

4.4 Are rent rates influenced by the student population in a college town? Let *rent* be the average monthly rent paid on rental units in a college town in the United States. Let *pop* denote the total city population, *avginc* the average city income, and *pctstu* the student population as a percent of the total population. One model to test for a relationship is

$$\log(rent) = \beta_0 + \beta_1\log(pop) + \beta_2\log(avginc) + \beta_3 pctstu + u.$$

(i) State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect.

(ii) What signs do you expect for β_1 and β_2?

(iii) The equation estimated using 1990 data from RENTAL.RAW for 64 college towns is

$$\widehat{\log(rent)} = \underset{(.844)}{.043} + \underset{(.039)}{.066}\log(pop) + \underset{(.081)}{.507}\log(avginc) + \underset{(.0017)}{.0056}\,pctstu$$
$$n = 64,\ R^2 = .458.$$

What is wrong with the statement: "A 10% increase in population is associated with about a 6.6% increase in rent"?

(iv) Test the hypothesis stated in part (i) at the 1% level.

4.5 Consider the estimated equation from Example 4.3, which can be used to study the effects of skipping class on college GPA:

$$\widehat{colGPA} = \underset{(0.33)}{1.39} + \underset{(.094)}{.412}\, hsGPA + \underset{(.011)}{.015}\, ACT - \underset{(.026)}{.083}\, skipped$$
$$n = 141,\ R^2 = .234.$$

(i) Using the standard normal approximation, find the 95% confidence interval for β_{hsGPA}.

(ii) Can you reject the hypothesis H_0: $\beta_{hsGPA} = .4$ against the two-sided alternative at the 5% level?

(iii) Can you reject the hypothesis H_0: $\beta_{hsGPA} = 1$ against the two-sided alternative at the 5% level?

4.6 In Section 4.5, we used as an example testing the rationality of assessments of housing prices. There, we used a log-log model in *price* and *assess* [see equation (4.47)]. Here, we use a level-level formulation.

(i) In the simple regression model

$$price = \beta_0 + \beta_1 assess + u,$$

the assessment is rational if $\beta_1 = 1$ and $\beta_0 = 0$. The estimated equation is

$$\widehat{price} = \underset{(16.27)}{-14.47} + \underset{(.049)}{.976}\, assess$$
$$n = 88,\ SSR = 165{,}644.51,\ R^2 = .820.$$

First, test the hypothesis that H_0: $\beta_0 = 0$ against the two-sided alternative. Then, test H_0: $\beta_1 = 1$ against the two-sided alternative. What do you conclude?

(ii) To test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, we need the SSR in the restricted model. This amounts to computing $\sum_{i=1}^{n} (price_i - assess_i)^2$, where $n = 88$, since the residuals in the restricted model are just $price_i - assess_i$. (No estimation is needed for the restricted model because both parameters are specified under H_0.) This turns out to yield $SSR = 209{,}448.99$. Carry out the F test for the joint hypothesis.

(iii) Now test H_0: $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$ in the model

$$price = \beta_0 + \beta_1 assess + \beta_2 sqrft + \beta_3 lotsize + \beta_4 bdrms + u.$$

The R-squared from estimating this model using the same 88 houses is .829.

(iv) If the variance of *price* changes with *assess*, *sqrft*, *lotsize*, or *bdrms*, what can you say about the F test from part (iii)?

4.7 In Example 4.7, we used data on Michigan manufacturing firms to estimate the relationship between the scrap rate and other firm characteristics. We now look at this example more closely and use a larger sample of firms.

(i) The population model estimated in Example 4.7 can be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

Using the 43 observations available for 1987, the estimated equation is

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\, \log(sales) + \underset{(.360)}{.992}\, \log(employ)$$
$$n = 43,\ R^2 = .310.$$

Compare this equation to that estimated using only 30 firms in the sample.

(ii) Show that the population model can also be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,$$

where $\theta_3 \equiv \beta_2 + \beta_3$. [Hint: Recall that $\log(x_2/x_3) = \log(x_2) - \log(x_3)$.] Interpret the hypothesis H_0: $\theta_3 = 0$.

(iii) When the equation from part (ii) is estimated, we obtain

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\, \log(sales/employ) + \underset{(.205)}{.041}\, \log(employ)$$
$$n = 43,\ R^2 = .310.$$

Controlling for worker training and for the sales-to-employee ratio, do bigger firms have larger statistically significant scrap rates?

(iv) Test the hypothesis that a 1% increase in *sales/employ* is associated with a 1% drop in the scrap rate.

4.8 Consider the multiple regression model with three independent variables, under the classical linear model assumptions MLR.1 through MLR.6:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You would like to test the null hypothesis H_0: $\beta_1 - 3\beta_2 = 1$.

(i) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS estimators of β_1 and β_2. Find $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ in terms of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the covariance between them. What is the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Write the t statistic for testing H_0: $\beta_1 - 3\beta_2 = 1$.

(iii) Define $\theta_1 = \beta_1 - 3\beta_2$ and $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Write a regression equation involving β_0, θ_1, β_2, and β_3 that allows you to directly obtain $\hat{\theta}_1$ and its standard error.

4.9 In Problem 3.3, we estimated the equation

$$\widehat{sleep} = \underset{(112.28)}{3{,}638.25} - \underset{(.017)}{.148}\, totwrk - \underset{(5.88)}{11.13}\, educ + \underset{(1.45)}{2.20}\, age$$
$$n = 706,\ R^2 = .113,$$

where we now report standard errors along with the estimates.

(i) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

(ii) Dropping *educ* and *age* from the equation gives

$$\widehat{sleep} = 3{,}586.38 - .151\ totwrk$$
$$(38.91)\quad (.017)$$
$$n = 706,\ R^2 = .103.$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

(iii) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?

(iv) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

4.10 Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. For concreteness, let *return* be the total return from holding a firm's stock over the four-year period from the end of 1990 to the end of 1994. The *efficient markets hypothesis* says that these returns should not be systematically related to information known in 1990. If firm characteristics known at the beginning of the period help to predict stock returns, then we could use this information in choosing stocks.

For 1990, let *dkr* be a firm's debt to capital ratio, let *eps* denote the earnings per share, let (log)*netinc* denote net income, and let (log)*salary* denote total compensation for the CEO.

(i) Using the data in RETURN.RAW, the following equation was estimated:

$$\widehat{return} = 40.44 + .952\ dkr + .472\ eps - .025\ netinc + .003\ salary$$
$$(29.30)\quad (.854)\quad (.332)\quad (.020)\quad (.009)$$
$$n = 142,\ R^2 = .0285.$$

Test whether the explanatory variables are jointly significant at the 5% level. Is any explanatory variable individually significant?

(ii) Now reestimate the model using the log form for *netinc* and *salary*:

$$\widehat{return} = -69.12 + 1.056\ dkr + .586\ eps - 31.18\ netinc + 39.26\ salary$$
$$(164.66)\quad (.847)\quad (.336)\quad (14.16)\quad (26.40)$$
$$n = 142,\ R^2 = .0531.$$

Do any of your conclusions from part (i) change?

(iii) Overall, is the evidence for predictability of stock returns strong or weak?

4.11 The following table was created using the data in CEOSAL2.RAW:

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	———	.112 (.050)	.100 (.049)
profmarg	———	−.0023 (.0022)	−.0022 (.0021)
ceoten	———	———	.0171 (.0055)
comten	———	———	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observations *R*-Squared	177 .281	177 .304	177 .353

The variable *mktval* is market value of the firm, *profmarg* is profit as a percentage of sales, *ceoten* is years as CEO with the current company, and *comten* is total years with the company.

(i) Comment on the effect of *profmarg* on CEO salary.

(ii) Does market value have a significant effect? Explain.

(iii) Interpret the coefficients on *ceoten* and *comten*. Are the variables statistically significant? What do you make of the fact that longer tenure with the company, holding the other factors fixed, is associated with a lower salary?

COMPUTER EXERCISES

4.12 The following model can be used to study whether campaign expenditures affect election outcomes:

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u,$$

where *voteA* is the percent of the vote received by Candidate A, *expendA* and *expendB* are campaign expenditures by Candidates A and B, and *prtystrA* is a measure of party

strength for Candidate A (the percent of the most recent presidential vote that went to A's party).

(i) What is the interpretation of β_1?

(ii) In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.

(iii) Estimate the model above using the data in VOTE1.RAW and report the results in usual form. Do A's expenditures affect the outcome? What about B's expenditures? Can you use these results to test the hypothesis in part (ii)?

(iv) Estimate a model that directly gives the *t* statistic for testing the hypothesis in part (ii). What do you conclude? (Use a two-sided alternative.)

4.13 Use the data in LAWSCH85.RAW for this exercise.

(i) Using the same model as Problem 3.4, state and test the null hypothesis that the rank of law schools has no ceteris paribus effect on median starting salary.

(ii) Are features of the incoming class of students—namely, *LSAT* and *GPA*—individually or jointly significant for explaining *salary*?

(iii) Test whether the size of the entering class (*clsize*) or the size of the faculty (*faculty*) need to be added to this equation; carry out a single test. (Be careful to account for missing data on *clsize* and *faculty*.)

(iv) What factors might influence the rank of the law school that are not included in the salary regression?

4.14 Refer to Problem 3.14. Now, use the log of the housing price as the dependent variable:

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u.$$

(i) You are interested in estimating and obtaining a confidence interval for the percentage change in *price* when a 150-square-foot bedroom is added to a house. In decimal form, this is $\theta_1 = 150\beta_1 + \beta_2$. Use the data in HPRICE1.RAW to estimate θ_1.

(ii) Write β_2 in terms of θ_1 and β_1 and plug this into the log(*price*) equation.

(iii) Use part (ii) to obtain a standard error for $\hat{\theta}_1$ and use this standard error to construct a 95% confidence interval.

4.15 In Example 4.9, the restricted version of the model can be estimated using all 1,388 observations in the sample. Compute the *R*-squared from the regression of *bwght* on *cigs*, *parity*, and *faminc* using all observations. Compare this to the *R*-squared reported for the restricted model in Example 4.9.

4.16 Use the data in MLB1.RAW for this exercise.

(i) Use the model estimated in equation (4.31) and drop the variable *rbisyr*. What happens to the statistical significance of *hrunsyr*? What about the size of the coefficient on *hrunsyr*?

(ii) Add the variables *runsyr*, *fldperc*, and *sbasesyr* to the model from part (i). Which of these factors are individually significant?

(iii) In the model from part (ii), test the joint significance of *bavg*, *fldperc*, and *sbasesyr*.

4.17 Use the data in WAGE2.RAW for this exercise.

(i) Consider the standard wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

State the null hypothesis that another year of general workforce experience has the same effect on log(*wage*) as another year of tenure with the current employer.

(ii) Test the null hypothesis in part (i) against a two-sided alternative, at the 5% significance level, by constructing a 95% confidence interval. What do you conclude?